Two-sample Ellipsoidal Bounding in the Context of Parameter Estimation

Garry Phillip Hollier

PhD Thesis

School of Electronic and Electrical Engineering,
The University of Birmingham,
November 1999

Synopsis

This thesis deals with the Fogel-Huang algorithm for the bounding by an ellipsoid of the intersection of an ellipsoid and a strip bounded by parallel hyperplanes, as applied to the task of parameter estimation. Although the Fogel-Huang ellipsoid is the unique minimum-value (Behrend-Löwner/John — BLJ) ellipsoid containing this intersection, the ellipsoid resulting from an iteration of the Fogel-Huang algorithm to find an ellipsoid bounding the intersection of an ellipsoid and several such strips is *not* the BLJ ellipsoid for that intersection.

After introductory material and a survey of the literature, the performance of the Fogel-Huang algorithm is examined, both in simulation and theoretically. This examination is made in terms of a characteristic length and the volume of the resulting ellipsoid, and comparisons are made with the BLJ ellipsoid containing the intersection of the strips.

In addition, the results of recycling the Fogel-Huang algorithm, with the hyperplane order fixed and varying, are examined in the same terms.

Two modifications of the Fogel-Huang algorithm are then proposed and examined. In the first, a family of ellipsoids containing the intersection of an ellipsoid and two strips is derived. The minimum-volume member of this family (not, in general, the BLJ ellipsoid for the intersection being bounded) is found.

In the second, a set of up to p (the dimension of the parameter space) strips is approximated by a more convenient set of strips whose intersection with the original ellipsoid contains the intersection of the original ellipsoid with the original strips. Then a family of ellipsoids containing the intersection of the original ellipsoid and the new strips is found. The volume-optimal member of this family is then determined.

The ellipsoids resulting from these modifications often have smaller volumes than the volume found by applying the Fogel-Huang algorithm, and so lead to closer approximations to the BLJ ellipsoid.

The fact that the first of these modifications leads to smaller bounding ellipsoids than the iterated Fogel-Huang algorithm motivates the investigation of methods resulting in the true BLJ ellipsoid for the intersection of an ellipsoid and two strips.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Parameterised Modelling

Many physical, industrial and biological processes can be modelled using state variables: in continuous time, $\dot{x} = f(x, u, t)$, $y = g(x, t)$, where x is a vector of variables describing the state of the process or system, u is a vector of inputs, and y is a vector of outputs; and in discrete time, $x_k = f_k(x_{k-1}, u_k)$, $y_k = g_k(x_k)$. The functions f, g, f_k and g_k encapsulate an understanding of the system, and, as this understanding is usually imperfect, the functions themselves will only be partially known.

If the number of state variables x is appropriate for the adequate explanation of the system, this uncertainty is often representable in the form of a finite number of unknown parameters, so that it is possible to write $f(x, u, t) = \tilde{f}(\theta; x, u, t)$, $g(x, t) = \tilde{g}(\psi; x, t)$, $f_k(x_{k-1}, u) = \tilde{f}_k(\theta; x_{k-1}, u)$, $g_k(x_k) = \tilde{g}_k(\psi; x_k)$ where the functions on the right-hand side are known and the task of improving knowledge of the functions f and g, or f_k and g_k, becomes that of better specifying the statistics of the parameters θ and ψ.

When this is done, the results can provide a check on the physical understanding which produced the model in the first place and allow better control over the system through the manipulation of those of the inputs u which are under human control.

In the absence of noise, or when noise is negligible, θ and ψ can be found by making a sufficient number (r, say — usually equal to the number of parameters) of observations of the input and output of the system and inverting the equations $\dot{x}(t_i) = f(x(t_i), u(t_i), t_i)$ and $y(t_i) = g(x(t_i), t_i)$, or $x_{k_i} = f_{k_i}(x_{k_i-1}, u_{k_i})$ and $y_{k_i} = g_{k_i}(x_{k_i})$, with $i = 1, \ldots, r$, in terms of θ and ψ (which is not always a trivial task).

In the more common situation where noise is present, a precise determination of θ and ψ is rarely possible; instead, observations can serve to replace, at each step, an *a priori* knowledge of the statistics (in a wide sense) of the distributions of θ and ψ by an *a posteriori* knowledge which is at least not less "certain". This requires known or assumed characteristics of the noise and possibly of the distribution of θ and ψ before the first observation.
Two methods which are often utilised are to assume

1. that the noise and initial *a priori* estimated distributions of θ and ψ have a Gaussian distribution with known mean and variance;

2. that the noise falls within a known, bounded region.

In the case where the models are linear in the parameters θ and ψ, approach 1 leads (through generalised least squares, see e.g. Norton [15]) leads to *a posteriori* Gaussian distributions of θ and ψ, where greater "certainty" with respect to the *a priori* distribution is manifested by a decrease in variance, and approach 2 leads to θ and ψ falling in an *a posteriori* region contained in the corresponding *a priori* region.
Approach 1 has the advantage of producing a more detailed description of the distribution of the parameters, but it also requires more detail to start with (the distribution of noise and the initial *a priori* of the parameters). Moreover, approach 2 provides a guaranteed set within which the parameters must lie, although it supplies no information regarding the probability distribution of the parameters within that set.

1.2 Bounded Parameter Estimation for Models Linear in Parameters

If the equations for the system can be summarised in the form

$$y_k = {n_k}^{\mathrm{T}}\theta + v_k,$$

where k is possibly indexes the components of the output vector as well as time, and n_k consists of known quantities, the model is linear in the fixed parameter vector $\theta \in \mathbb{R}^p$. If the error v_k is bounded, say $e_k^m \leq v_k \leq e_k^M$, where e_k^m and e_k^M are known scalars, then, by rescaling the equation (see page 25), e_k^m and e_k^M can be set to -1 and 1 respectively. Then, for each k, θ is restricted to the set $\Pi_k = \{\theta \in \mathbb{R}^p : y_k - 1 \leq {n_k}^{\mathrm{T}}\theta \leq y_k + 1\}$ lying between the hyperplanes

$\mathbb{H}_k^{\pm} = \{\theta \in \mathbb{R}^p : {n_k}^{\mathrm{T}}\theta - y_k = \pm 1\}$. If none of the regression vectors n_k are parallel, the set $\Pi_1 \cap \ldots \cap \Pi_k$, the posterior feasible parameter set after k observations, will be a bounded polytope if $k \geq p$. Of course, the posterior feasible parameter set can be described by enumerating the vertices of the polytope and listing which vertices are neighbours, but such a description may be both complex and computationally expensive to obtain. An alternative description is in terms of simple sets which are guaranteed to contain the feasible parameter set, such as ellipsoids. The basic problem of this thesis is to find ellipsoids containing this polytope.

1.3 Ellipsoidal Methods for Bounded Parameter Estimation

Suppose the initial ellipsoid is given by $\mathcal{E}_0 = \mathcal{E}(a_0, Q_0) = \{\theta \in \mathbb{R}^p : (\theta - a)^{\mathrm{T}} Q_0^{-1} (\theta - a) \leq 1\}$ in which the true value of the parameter vector θ_t is guaranteed to lie. If an observation is made which implies that the true value lies in $\Pi_1 = \Pi(n_1, y_1) = \{\theta : ({n_1}^{\mathrm{T}}\theta - y_1) \leq 1\}$, then, for all ellipsoids $\mathcal{E}$ such that $\mathcal{E} \supset \mathcal{E}_0 \cap \Pi_1$, $\theta_t \in \mathcal{E}$, so all such ellipsoids are estimates for θ_t. The idea of the ellipsoidal methods for bounded parameter estimation is to find a family of ellipsoids containing $\mathcal{E}_0 \cap \Pi_1$ and to find an ellipsoid in that family, $\mathcal{E}_1$, say, which minimises some criterion measuring the size of the ellipsoids of the family over the family. This ellipsoid becomes the new *a priori* set for the new observation.

1.4 The New Methods in this Thesis

There are three main new methods in this Thesis. The first finds the minimum-volume ellipsoid in a family containing the intersection of two strips and a prior ellipsoid. This "minimum-volume ellipsoid in a family" is not necessarily *the* minimum-volume ellipsoid containing the intersection. The second looks at the intersection of a set of s strips and an ellipsoid, replacing $s - 1$ of the strips by more convenient strips and finding the minimum-volume ellipsoid in a family containing the intersection of the replacement strips and the original ellipsoid. The last method is intended to find the actual minimum-volume ellipsoid about the intersection of two strips and an ellipsoid.

Chapter 2

Literature Survey

2.1 Introduction

Given a system with state equations

$$\begin{aligned} x_k &= A_k x_{k-1} + B_k u_k + v_k \\ y_k &= C_k x_k + w_k, \end{aligned}$$

where $x_k, v_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^\ell$, $y_k, v_k \in \mathbb{R}^s$, $A_k \in \mathbb{R}^{n\times n}$, $B_k \in \mathbb{R}^{n\times \ell}$, $C_k \in \mathbb{R}^{s\times n}$, x_k is the state, u_k is the input, y_k is the observed output and v_k and w_k are noise, state estimation attempts to maximise the knowledge of the state x_k, given the parameters A_k, B_k and C_k, the inputs u_k, and partial knowledge of v_k and w_k, either in the form of (narrowly) statistical information such as their mean and covariance, or in the form of bounds. The major difference between state and parameter estimation is that the former requires an update between observations, as the state is not fixed, so the state estimate must be propagated using the system equations, whereas this update is not required for static parameters (actually, many parameter estimation schemes incorporate the possibility of drift in the parameters, but this will not be dealt with here).

For example, if it is known that $x_0 \in \mathcal{X}_0 = \mathcal{X}_0^+ \subset \mathbb{R}^n$, $v_k \in \mathcal{V}_k \subset \mathbb{R}^n$, $w_k \in \mathcal{W}_k \subset \mathbb{R}^s$, then the state x_k obeys

$$\begin{aligned} x_k \in \mathcal{X}_k^+ &= \mathcal{X}_k^- \cap \{x : y_k - C_k x \in \mathcal{W}_k\} && \text{(observation)} \\ \subset \mathcal{X}_k^- &= A_k \mathcal{X}_{k-1}^+ + B_k u_k + \mathcal{V}_k. && \text{(propagation)} \end{aligned}$$

For parameter estimation, if the model can be written

$$y_k = \Phi_k \theta + v_k,$$

where θ has an initial *a priori* covariance P_0, Φ_k is a known matrix of explanatory variables and v_k is a zero-mean random variable with covariance R_k, one statistical method is *minimum-covariance linear unbiased updating* Norton [15]. It is proposed that an estimate $\hat{\theta}_k$ for θ at time k should be given by $\hat{\theta}_k = J_k\hat{\theta}_{k-1} + K_k y_k$, where J_k and K_k are matrices of the appropriate dimensions and are functionally independent of θ_{k-1} and y_k, i.e., the new estimate is a linear function of the old estimate and the new observation.

Insisting that the estimate should be unbiased yields

$$\theta = \mathrm{E}\hat{\theta}_k = J_k\mathrm{E}\hat{\theta}_{k-1} + K_k\Phi_k\theta = (J_k + K_k\Phi_k)\theta,$$

where E is the expectation operator, or $J_k + K_k\Phi_k = I$.

Then the covariance of $\hat{\theta}_k$ is

$$\begin{aligned}
P_k &= \mathrm{E}(\hat{\theta}_k - \mathrm{E}\hat{\theta}_k)(\hat{\theta}_k - \mathrm{E}\hat{\theta}_k)^{\mathrm{T}} \\
&= \mathrm{E}(\hat{\theta}_k - \theta)(\hat{\theta}_k - \theta)^{\mathrm{T}} \\
&= \mathrm{E}((I - K_k\Phi_k)\hat{\theta}_{k-1} + K_k y_k - \theta)(\hat{\theta}_k - \theta)^{\mathrm{T}} \\
&= \mathrm{E}((I - K_k\Phi_k)(\hat{\theta}_{k-1} - \theta) + K_k(y_k - \Phi_k\theta))(\hat{\theta}_k - \theta)^{\mathrm{T}} \\
&= (I - K_k\Phi_k)P_{k-1}(I - K_k\Phi_k)^{\mathrm{T}} + K_k R_k {K_k}^{\mathrm{T}},
\end{aligned}$$

on the assumption that $\mathrm{E}(\hat{\theta}_{k-1} - \theta)v_k = 0$.

Then $K_k - P_{k-1}{\Phi_k}^{\mathrm{T}}(\Phi_k P_{k-1}{\Phi_k}^{\mathrm{T}} + R_k)^{-1}$ minimises this P_k (as can be seen by incrementing K_k to $K_k + \delta K_k$ and setting the corresponding increment δP_k in P_k to zero), with minimum value $P_k = (I - K_k\Phi_k)P_{k-1}$. This equation is used to update the covariance after each new data item, and the minimising K_k and the J_k related to it by $J_k = I - K_k\Phi_k$ are used to update x_k according to the linear formula above.

There is a corresponding method in state estimation, *Kalman filtering*, which can be derived in a similar fashion (see Grewal and Andrews [11]).

2.2 Statistical Methods for Parameter Estimation

Pronzato and Walter[21] compare the statistical approach of D-optimal design, which maximises the determinant of the Fisher information matrix[1] for the maximum likelihood estimate of a parameter vector, with what the authors dub V-optimal design, in a bounded error context. V-optimal design seeks to minimise the volume of the posterior feasible set in parameter space. When the number of experiments (the word *experiment* is used for an observation of the output when it is wished to emphasise that some, at least, of the inputs are chosen with the aim of obtaining information about the system) is equal to the number of parameters, the V-optimal and D-optimal approaches result in the same experiments, but when the number of experiments exceeds the number of parameters, a complication arises: the volume of the feasible set depends on the outcome of the experiments (in the case of one experiment per parameter, only the location, but not the shape or volume, of the feasible set, depends on the experimental outcome). The authors' solution to this problem is to replace V-optimality by $\hat{V}$-optimality, where the volume of an estimated feasible set $\hat{S}$ is minimised instead of the volume of the feasible set S itself.

In the $\hat{V}$-optimality approach, a nominal value for the parameter vector is required (as also the case in the D-optimality method when the model is not linear in the parameters), which, as the authors acknowledge, is a weakness.

In many cases, D-optimal experiments are repeated when the number of allowable experiments exceeds the number of parameters to be identified, but repeated experiments do not lead to a reduction of the estimate feasible set $\hat{S}$, so a series of $\hat{V}$-optimal experiments contains no repetitions. Moreover, an $(N+1)$-member series of $\hat{V}$-optimal experiments will not, in general, contain an N-member series of $\hat{V}$-optimal experiments, so one cannot simply extend a run of such experiments until sufficient knowledge has been attained, without departing from $\hat{V}$-optimality.

On the other hand, the authors provide a simple criterion for determining whether there exists a finite series of $\hat{V}$-optimal experiments which reduces the volume of $\hat{S}$ to its infimum.

---

[1]For a probability density function $p_x(t)$ for random variables $x_1, \ldots, x_n$, the *Fisher information matrix* is given by

$$\begin{bmatrix} \frac{\partial^2}{\partial t_1^2} \ln p_x(t) & \frac{\partial^2}{\partial t_1 \partial t_2} \ln p_x(t) & \cdots \\ \frac{\partial^2}{\partial t_2 \partial t_1} \ln p_x(t) & \frac{\partial^2}{\partial t_2^2} \ln p_x(t) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

For a Gaussian probability density function, the Fisher information matrix is the inverse of the covariance.

Kurzhanski and Tanaka[18] (discussed in more detail in the next section) achieve a partial unification of the statistical and bounding approaches by considering Gaussian distributions of the error, where the variance is known, but the mean is only known to lie within a convex set. Then, the posterior bounds on the mean resulting from observations obey the same equations as the parameters in a pure bounding context.

2.3 Bounding

The paper just mentioned by Kurzhanski and Tanaka[18] provides a very general framework for bounding techniques by considering the system $y(k) = Cu_k + v_k$, where y is a sequence of measurements in $\mathbb{R}^m$, u a sequence of known inputs in $\mathbb{R}^m$, v a noise sequence and $C \in \mathbb{R}^{m\times n}$ is a matrix of unknown parameters. It is known that $C \in \mathbf{C}_0$, $v_k \in \mathbf{Q}_k$, where $\mathbf{C}_0$ and $\mathbf{Q}_k$ are compact and convex.

Through manipulation of support functions ρ ($\rho(x|\mathcal{S}) = \sup_{z\in\mathcal{S}}\{z^{\mathrm{T}}x\}$), the authors derive an expression for the feasible parameter set (called the *informational domain* by the authors), $\mathbf{C}[s]$, the set of C consistent with the sequence $y_k, k \in \{1,\ldots,s\}$, as the set of C such that

$$\bar{C} \in \cap_{M[1,s]\in\mathbb{R}^{mn\times m}[1,s]}\{(I_{mn} - \sum_{k=1}^{s}({u_k}^{\mathrm{T}} \otimes I_m))\bar{C}_0 + \sum_{k=1}^{s} M_k(y_k - \mathbf{Q}_k)\},$$

where $\mathbb{R}^{mn\times m}[1,s]$ is the set of sequences $A(1),\ldots,A(s)$ in $\mathbb{R}^{mn\times m}$, and, if $A \in \mathbb{R}^{p\times q}$, $\bar{A} \in \mathbb{R}^{pq}$ is obtained by stacking the columns of A. Also, $\otimes$ denotes the Kronecker matrix product:

$$A \otimes B = [a_{ij}] \otimes [b_{ij}] = \begin{bmatrix} a_{11}B & a_{12}B & \ldots \\ a_{21}B & a_{22}B & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

If the sequence $\mathbf{C}[k], k = 1,\ldots s$ is imagined as a series of subsets of the hyperplanes $H_k\{x = (C,\xi) \in \mathbb{R}^{m\times n}\times\mathbb{R} : \xi = k\}$, and the boundaries of the $\mathbf{C}[k] \cap H_k$ are thought of as being joined up, with each point $x_{k-1} \in \partial\mathbf{C}[k-1] \cap H_{k-1}$ connected by a line segment to a point $f_{k-1,k}(x_{k-1}) \in \partial\mathbf{C}[k]\cap H_k$, where $fk-1,k$ is a homeomorphism $\partial\mathbf{C}[k-1]\cap H_{k-1} \mapsto \partial\mathbf{C}[k]\cap H_k$, then the resulting set is a "funnel" in $\mathbb{R}^{m\times n}\times\mathbb{R}$ (the situation in the continuous counterpart to this is more easily imagined!).

Now, obviously,

$$\mathbf{C}[s] \subseteq (I_{mn} - \sum_{k=1}^{s} M_k({u_k}^{\mathrm{T}} \otimes I_m)\bar{C}_0 + \sum_{k=1}^{s} M_k(y_k - \mathbf{Q}_k)$$

for any particular sequence $M[1,s]$ in $\mathbb{R}^{mn\times m}$.

The authors then conclude that schemes for bounding $\mathbf{C}[s]$ corresponding to different sequences $M[1,s]$ lend themselves to parallel computation, where each independent processor utilises a different sequence. Obviously, such an approach would be more expensive than using just one processor, but would rapidly deliver a great deal of information on the feasible set at each stage, as each processor would provide a different bounding set. However, the authors do not address the difficulties of integrating the information from each processor quickly into a simple format, such as a simple approximant to the intersection of the various bounding sets.

Kostousova and Kurzhanski[17] specialise the ideas of the previous paper to deal with polytope bounding for discrete time and ellipsoidal bounding for continuous time, investigating attainability domains (the set of states consistent with initial conditions and input sequences in given sets, called the *reachable set* elsewhere in the literature) in both cases and state-estimation for the continuous estimation technique.

Veres and Norton[24] discuss errors-in-variables models, showing how an apparently linear model can lead to nonlinear bounds for the feasible parameter set. In such a model, the output y_k at time k is given by $y_k = f(\phi_k, \theta) + e_k$, where ϕ_k is a vector of explanatory variables and e_k is the error due to mis-specification of the model structure. The variables ϕ_k are only known with uncertainty. In this paper, this uncertainty is given in the form of a bound on the error $\tilde{\phi}_k := \phi_k - \phi_k^o$ between the variable and the observation corresponding to it: $|\tilde{\phi}_k^i| \leq \epsilon_\phi^i, i = 1, \ldots, q$. Bounds are also given for $\tilde{x}_k := x_k - x_k^o$ and e_k: $|\tilde{x}_k| \leq \epsilon_y$, $|e_k| \leq \epsilon_k$. For models linear in both the explanatory variables, ϕ, and the parameters, θ, $f(\phi_k, \theta) = F(\phi_k)\theta$, where F is a matrix, the authors make the distinction between *static* and *dynamic* models.

Let $\phi_k' = [y_k\ {\phi_k}^{\mathrm{T}}]^{\mathrm{T}}$, $\theta' = [-1\ \theta^{\mathrm{T}}]^{\mathrm{T}}$ and define $F'(\phi_k')$ so that $y_k = F(\phi_k)\theta + e_k$ is equivalent to $F'(\phi_k')\theta' + e_k = 0$. Then a static model has successive vectors ϕ_k' deterministically unrelated, whereas a dynamic model has variables appearing in successive vectors ϕ_k'.

In a static model, the bounds on the parameters resulting from observations are given by a sum of polytopes. in a dynamic model, one can ignore the interactions between bounds (in effect, treating multiple occurrences of explanatory variables in the model equations at different times as the occurrence of different variables with the same observed values) which may result in much looser (but simpler) bounds on the parameters, or these interactions can be taken into account in finding the feasible parameter set. Then, the boundary of this feasible parameter

set may be nonlinear.

It is also shown that the result of a single observation is a pair of hyperplane bounds for each orthant of parameter space which has a nonempty intersection with the resulting feasible parameter set.

In Walter and Piet-Lahanier[25], the authors discuss the shortcomings of statistical approaches; primarily the fact that the assumptions about probability density functions for the noise may not be well founded, and the amount of data might not be sufficient to check these assumptions. They also criticise the use of noise to model structural error for reduced-order systems or nonlinear ones (although bounding techniques do not require assumptions about the mean or correlation of noise, which might be hard to justify, it is difficult to see why bounds would be better than statistics in practice, in general).

For models linear in parameters, the authors deal with ellipsoidal bounding, axis-parallel orthotope bounding and the exact description of the posterior feasible set.

As ellipsoidal bounding as covered in this paper is dealt with in Section 3.1.1 of Chapter 3 below, orthotopic bounding will be examined next. The idea behind this is to find the maximum and minimum values of the coordinates consistent with the inequalities derived from the observations (so it is more properly axis-aligned orthotopic bounding). In p-dimensional parameter space with N data points, the orthotope bounds can be found "all at once" found by solving $2p$ linear programming problems, each with $2N$ inequalities. However, using some ellipsoidal bounding technique as a preprocessor enables some redundant inequalities to be removed, thus reducing the problem to one of lesser complexity. Nevertheless, the worst-case complexity is still great if p or N is large.

Moreover, singling out the directions given by the axes for special treatment may result in poor performance if the polytope given by exact solution of the bounds is not close to being axis-aligned.

Orthotopic bounding can also be implemented recursively, reducing the computational complexity but increasing the volume of the resulting orthotope. An initial prior orthotope must also be supplied.

Exact methods derive the polytope bounding the feasible parameter set. This polytope can be represented in a variety of ways, from a list of all j-faces ($j = 0, \ldots, p-1$) with each entry associated with a list of all $(j-1)$-faces it contains and, in the case of $j = 0$ or 1, a co-ordinate geometrical description of the j-face itself, to a list of all vertices with a list of their neighbours

and one of their supporting hyperplanes associated with each entry. The algorithms relying on the latter type of representation are recursive, and with each new inequality, they discard any vertex not satisfying the inequality, and add the vertices defined by the intersection of the hyperplane associated with the inequality and any edge running between a discarded vertex and an adjacent retained vertex.

The drawbacks of the exact methods are that they are computationally intensive compared to the ellipsoidal bounding schemes and the resulting polytope may be extremely complicated (although, the Walter and Piet-Lahanier state, they are often not, due to the large number of redundant inequalities in the typical case).

Much of the same ground is also covered in Pronzato and Walter[23].

2.4 Ellipsoidal Bounding

Apart from what is covered below, Pronzato and Walter[23, 22], address the fact that although the Fogel-Huang algorithm with the Belforte-Bona-Cerone modification results in an ellipsoid which has the minimum volume of all those which contain the intersection of a previous ellipsoid and the strip between the current pair of parallel hyperplanes, but this does not mean that applying it sequentially to the resulting ellipsoids and new strips results in the volume-optimal ellipsoid containing the intersection of all the strips and the original ellipsoid. However, this volume-optimal ellipsoid will be the volume-optimal ellipsoid containing the vertices of the polytope defined by the intersection of the strips if this intersection is contained in the original ellipsoid, so that further intersecting the intersection of the strips with the ellipsoid results in no reduction of the former set. This is so, because there exists a unique minimum-volume ellipsoid containing a given convex set (see the discussion of the next paper) and a polytope is convex. But there exists an algorithm which can find the volume optimal ellipsoid containing a set of points with full-dimensional convex hull.

The authors derive this algorithm by using Lagrangian techniques and ideas from experimental design. However, the convergence of the algorithm is very slow.

In the context of improving Khachian's algorithm for solving systems of linear inequalities with polynomial complexity in the data, König and Pallaschke[16] show that the Behrend-Löwner/John ellipsoid (i.e., the unique minimum volume ellipsoid containing a given convex compact set. The existence and uniqueness of this ellipsoid was attributed to Behrend and Löwner by Berger [2] and to John and Löwner by Grötschel *et al.* [12]. Accordingly, this

minimum-volume ellipsoid is denoted the *BLJ* ellipsoid.) for the ellipsoidal section defined by

$$S = \left\{ x \in \mathbb{R}^p : (x-a)^{\mathrm{T}} Q^{-1} (x-a) \leq 1, \frac{n^{\mathrm{T}}(x-a)}{\sqrt{n^{\mathrm{T}} Q n}} \in [\xi, \eta] \right\},$$

where $Q > 0$, $-1 \leq \xi < \eta \leq 1$, is given by

$$\mathcal{E} = \{ x \in \mathbb{R}^p : (x-\bar{a})^{\mathrm{T}} \bar{Q}^{-1} (x-\bar{a}) \leq 1 \},$$

where

$$\begin{aligned} \bar{a} &= a - \gamma z, \\ \bar{Q} &= \beta^2 \left(Q - \left(1 - \left(\frac{\alpha}{\beta} \right)^2 \right) z z^{\mathrm{T}} \right) \end{aligned}$$

with

$$\begin{aligned} z &= \frac{1}{\sqrt{a^{\mathrm{T}} Q a}} Q a \\ \beta^2 &= \frac{p^2}{p^2 - 1} \left[1 - \frac{\eta^2 + \xi^2}{2} + \sqrt{\left(\frac{\eta^2 - \xi^2}{2} \right)^2 + \frac{(1-\eta^2)(1-\xi^2)}{p^2}} \right] \\ \alpha &= (\eta - \xi) \left[\sqrt{1 - \frac{1-\eta^2}{\beta^2}} + \sqrt{1 - \frac{1-\xi^2}{\beta^2}} \right]^{-1} \\ \gamma &= \xi + \alpha \sqrt{1 - \frac{1-\xi^2}{\beta^2}}. \end{aligned}$$

They do this by making the affine coordinate transformation which takes the ellipsoid defining the section into the unit sphere and the normal to the hyperplanes defining the section to be along a coordinate axis. The symmetry of the geometry about this axis is exploited to find the minimum volume ellipsoid in the image coordinates, and then the fact that affine transformations preserve ratios of volumes is used to deduce that the minimum volume ellipsoid in the original coordinates is the inverse image of the minimum volume ellipsoid in the transformed coordinates.

They further show that the ratio of the volumes of $\mathcal{E}$ and $\{ x \in \mathbb{R} : (x-a)^{\mathrm{T}} Q^{-1} (x-a) \leq 1 \}$ is

$$(\eta - \xi) \frac{\beta^p}{\sqrt{\beta^2 - (1-\eta^2)} + \sqrt{\beta^2 - (1-\xi^2)}}.$$

On making the substitutions $\xi = (\nu - 1)/\sqrt{g}$, $\eta = (\nu + 1)/\sqrt{g}$ to give ξ and η in terms of quantities employed later, it can be seen that the above expression for the volume ratio is equal

to that given below for the Fogel-Huang ellipsoid (equation (3.16)), and so the Fogel-Huang ellipsoid (as modified by Belforte, Bona and Cerone) is the BLJ ellipsoid.

Durieu *et al.*[6, 7], consider ellipsoidal state bounding where there is a prediction step involving the approximation of an algebraic sum ($A + B = \{a + b : a \in A, b \in B\}$) of ellipsoids, alternating with a correction step involving the approximation of the intersection of ellipsoids. In both cases, the approximation consists in finding a family of ellipsoids each member of which contains what is being approximated, and then finding the member of the family with minimum size according to some criterion. The two criteria considered are trace, corresponding to the sum of squares of the lengths of the semi-axes of the ellipsoid, and determinant, corresponding to the product of the semi-axis lengths.

This work show some of the advantages that the trace criterion has over the determinant criterion. Firstly, for general ellipsoids, an explicit solution minimising the trace can be found for the prediction step (involving the approximation of sums), whereas no explicit solution is available for the determinant criterion.

Secondly, when the sum of K ellipsoids is being approximated, an approximation can be derived by taking all the K ellipsoids at once and approximating their sum, or by setting the first ellipsoid to be the first current ellipsoid and successively approximating the sum of the kth ellipsoid and the $(k-1)$st current ellipsoid by the kth current ellipsoid, i.e., by making a recursive approximation. When the trace criterion is used, the "all-at-once" and recursive approximations are equal (in the limit of infinite numerical precision), but, when the determinant criterion is utilised, each step of the recursion introduces an error compared to the "all-at-once" approximation.

Thirdly, as the determinant criterion corresponds to minimising a product, some of the multiplicands for the optimal solution can be very large if others are very small, and so the optimal ellipsoid may be very long and thin, resulting in great uncertainty in some directions. Durieu *et al.* also demonstrate that this can happen in practice. However, this "long-thinness" can always be removed by rescaling the parameter space, so it is no problem unless there are particular reasons for preferring the given scaling.

In addition, if $\det Q = \Delta(\lambda_1, \ldots, \lambda_p)$ is treated as a function of the semi-axis lengths λ_i, and $\mathrm{Trace}Q = \Sigma(\lambda_1, \ldots, \lambda_p)$ is treated in the same way, the change in Δ resulting from a change in λ is proportional (in the limit of small $\delta\lambda$) to

$$\frac{\delta\lambda_1}{\lambda_1} + \cdots + \frac{\delta\lambda_p}{\lambda_p}$$

whereas the change in Σ is proportional to $\delta\lambda_1 + \ldots + \delta\lambda_p$, so the determinant criterion treats proportional changes in the semi-axis lengths equally, whereas the trace criterion treats absolute changes equally. This is a good reason for preferring the determinant criterion[2].

Filippova *et al.*[8] consider ellipsoid bounding in the framework of Kurzhanski and Tanaka[18]. They deal with ellipsoidal bounding as a tool in continuous-time state estimation. The reachable set and guaranteed state estimation (the problem of finding as small as possible a set in state space which is guaranteed to contain the actual state of the system) are investigated here.

What happens if the error bounds are erroneous, either too loose or too tight? Maksarov and Norton[19] address this problem, and also modify the hyperplane shifting adjustment of Belforte, Bona and Cerone, by moving the current hyperplanes at each recursive step to be tangent to the most distant of all the previous ellipsoids (if this reduces the set of included parameters!), rather than just to the current ellipsoid.

When "too many" data points are uninformative, and too loose noise bounds are believed to be the reason for this, the bounds are provisionally reduced in magnitude until the closest hyperplane of the relevant pair becomes tangent to the current ellipsoid.

On the other hand, if the bounds are too tight, the feasible parameter set may disappear. In this case, the authors relax the bounds so that the current strip intersects the current ellipsoid. This approach is incompatible with the detection of outliers (data members which, due to, for example, faulty measurements, violate valid bounds) if the bounds are expanded to include all data. However, this incompatibility can be removed if limits to the acceptable expansion of bounds are imposed.

2.5 Models Nonlinear in Parameters

In Jaulin and Walter[13] the problem to be solved is: given model output $y_m(\theta) \subset \mathbb{R}^n$ and data $y \subset \mathbb{R}^n$, find $S \subset P \subset \mathbb{R}^p$ such that $y = y_m(\theta) + v$, where $v \in \mathcal{V} \subset \mathbb{R}^n$, $\forall\theta \in S$. P is the prior feasible set, S is the posterior feasible set. The authors recast this in the form $S = y_m^{-1}(y - \mathcal{E}) = y_m^{-1}(\mathcal{Y})$, which is a problem of set inversion. The approach they use to solve this problem is interval analysis. They make the following definitions:

Interval: $[\theta] := [\theta^-, \theta^+]$, $\theta^- \leq \theta^+$;

Box: $[\theta] := [\theta_1^-, \theta_1^+] \times \cdots \times [\theta_n^-, \theta_n^+]$;

[2]Thanks are due to Professor J.P. Norton for making this point

the width of a box is $w([\theta]) = \max\{\theta_i^+ - \theta_i^- : [\theta] = [\theta_1^-, \theta_1^-] \times \cdots \times [\theta_n^-, \theta_n^+]\}$;

Set of Boxes: $\square\mathbb{R}^n = \{[\theta] = [\theta_1^-, \theta_1^+] \times \cdots \times [\theta_n^-, \theta_n^+] : (\theta_1^-, \ldots, \theta_n^-), (\theta_1^+, \ldots, \theta_n^+) \in \mathbb{R}^n\}$;

It is noted that $\mathbb{R}^n \subset \square\mathbb{R}^n$.

For a given set A, $[A] = \cap_{[\theta] \supset A}[\theta]$.

the minimal inclusion function for $f : \mathbb{R}^m \to \mathbb{R}^n$ is $[f] : \square\mathbb{R}^m \to \square\mathbb{R}^m : [\theta] \mapsto [f([\theta])]$;

an *inclusion function* for f, $\mathcal{F} : \square\mathbb{R}^m \to \square\mathbb{R}^n$ satisfies:

1. $[f]([\theta]) \subset \mathcal{F}([\theta])$;
2. $\max_i(\theta_i^+ - \theta_i^-) = w([\theta]) \to 0 \Rightarrow w(\mathcal{F}([\theta])) \to 0$.

a paving is a union of boxes;

the accuracy of a paving $\mathcal{K}$ is $\min_{[\theta] \in \mathcal{K}}\{w([\theta])\}$.

The procedure is to start with an initial prior box, $[\theta](0)$, decide on the required accuracy for paving, ϵ_r, then to replace, if necessary, $[y_m]$ by an inclusion function $\mathcal{F}$ for y_m and then find $\mathcal{K}_{\mathbf{in}}$, $\mathcal{K}_{\mathbf{out}}$ such that $\mathcal{K}_{\mathbf{in}} \subset \mathcal{X} \subset \mathcal{K}_{\mathbf{out}}$, where $\mathcal{X} = \mathcal{F}^{-1}(\mathcal{Y})$ and $\mathcal{K}_{\mathbf{in}}$, $\mathcal{K}_{\mathbf{out}}$ are pavings. If the accuracy of $\mathcal{K}_{\mathbf{out}} - \mathcal{K}_{\mathbf{in}}$ is ϵ_r (the boxes in $\mathcal{K}_{\mathbf{in}}$, considered as a collection of boxes rather than their union, are also in $\mathcal{K}_{\mathbf{out}}$) then $\mathcal{K}_{\mathbf{in}}$, $\mathcal{K}_{\mathbf{out}} \to \mathcal{X}$ in an appropriate metric (which implies that $L_\infty(\mathcal{K}_{\mathbf{in}}, \mathcal{X})$, $L_\infty(\mathcal{K}_{\mathbf{out}}, \mathcal{X}) \to 0$) as $\epsilon_r \to 0$. This is achieved through the **S**et **I**nverter **V**ia

Interval **A**nalysis (SIVIA) algorithm:

Algorithm: 2.1

1. Initialisation:

   $k = 0$; stack $= \emptyset$; $\mathcal{K}_{\text{in}} = \emptyset$; $\mathcal{K}_{\text{i}} = \emptyset$ ($\mathcal{K}_{\text{i}}$ is the collection of *indeterminate* boxes such that neither $\mathcal{F}([\theta]) \subset \mathcal{Y}$ nor $\mathcal{F}([\theta]) \cap \mathcal{Y} = \emptyset$; $\mathcal{K}_{\text{out}} = \mathcal{K}_{\text{in}} \cup \mathcal{K}_{\text{i}}$).

2. Iteration:

   (a) If $\mathcal{F}([\theta](k)) \subset \mathcal{Y}$ then $\mathcal{K}_{\text{in}} := \mathcal{K}_{\text{in}} \cup [\theta](k)$. Go to 2d;

   (b) If $\mathcal{F}([\theta](k)) \cap \mathcal{Y} = \emptyset$ go to 2d;

   (c) If $w([\theta](k)) < \epsilon_r$, $\mathcal{K}_{\text{i}} := \mathcal{K}_{\text{in}} \cup [\theta](k)$;
   Else bisect $[\theta](k)$ along a principal plane (orthogonal to a maximal axis) and stack resulting boxes;

   (d) If the stack is nonempty, unstack into $[\theta](k+1)$, increment k and go to 2a;

In Jaulin and Walter[14], the authors modify SIVIA so that outliers do not result in an empty $\mathcal{K}_{\text{out}}$; rather, if j is the number of allowed outliers, $\mathcal{K}_{\text{out}}(j)$ will be the paving obtained when any box which violates the model for up to j observations (not necessarily the same observations for each box, which means that the resulting paving is a union of sets of boxes each of which violates the model for a subset $\mathcal{O}$ of the observations, where each $\mathcal{O}$ is such that $\overline{\overline{\mathcal{O}}} \leq j$) is retained.

Pronzato and Walter[22] show that errors in variables problems can be treated by considering models $y_m(k,\theta) = \phi_m(k,\theta)^{\mathrm{T}}\theta$, where the explanatory variables in the uncertain regressor ϕ_m may depend on θ. The error $\epsilon_\phi(k,\theta) = \phi_k - \phi_m(k,\theta)$ is bounded, and each pair of bounds on components of ϵ_ϕ translates into a set of 2^{p+1} linear bounds on θ, two of which apply in each orthant of p-dimensional parameter space. However, the pair of bounds for an orthant is not in general parallel, and so an algorithm finding the minimum volume ellipsoid containing the intersection of an ellipsoid and two half-spaces is required. The authors utilise the single-cut algorithm given in Grötschel *et al* [12] twice for this purpose.

Walter and Piet-Lahanier[25] discuss three general methods for dealing with models nonlinear in parameters: multiple linearisation, output error models and errors in variables (considered under the previous paper). Multiple linearisation linearises error bounds about a nominal value of the parameter vector θ, finds the minimal (respectively, maximal) values of the components

satisfying the linearised bounds, and then checks whether these minimal (respectively, maximal) values satisfy the original nonlinear bounds. If they do, an inner estimate of the parameter bounds is obtained. If they do not, an inner estimate can be found by bisecting the interval with end points given by the original nominal value and the failed bound.

In output-error models, the error is modified to contain the nonlinearity, and bounds on this redefined error are derived from the bounds on the original error, but these may be quite loose. In addition, the authors consider specific methods tailored to particular problem in hand: orthotopic bounding, which, as in the linear case, corresponds to the solution of $2N$ optimisation problems in $2p$ inequalities, this time through nonlinear programming techniques where these are available; scanning the parameter space either through a Monte Carlo technique or by utilising projections; interval analysis, considered elsewhere.

Chapter 3

The Fogel-Huang Algorithm

3.1 Introduction

3.1.1 Summary of "Minimum Volume Ellipsoids" (part) [22, Pronzato and Walter].

The Fogel-Huang algorithm finds the minimum-volume ellipsoid $\mathcal{E}_1$ containing the intersection $\mathcal{E}_0 \cap \Pi_1$ of an ellipsoid $\mathcal{E}_0$ and the region Π_1 bounded by a pair of parallel hyperplanes. (As this intersection is convex, there exists a unique minimum-volume ellipsoid containing it by the Behrend-Löwner/John theorem [2]). The algorithm repeats this process to obtain the BLJ ellipsoid $\mathcal{E}_n$ for the ellipsoid $\mathcal{E}_{n-1}$ and the hyperplane pair Π_n. If the Π_i are bounds on sets of parameters, for example, then the $\mathcal{E}_i$ are relatively simple sets guaranteed to contain the true parameter value. However, the Fogel-Huang algorithm suffers from a defect. Despite the fact that $\mathcal{E}_1$ obtained by using the Fogel-Huang algorithm is guaranteed to be the BLJ ellipsoid containing $\mathcal{E}_0 \cap \Pi_1$, $\mathcal{E}_2$ is not necessarily the minimum-volume ellipsoid containing $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, although, of course, it is the minimum-volume ellipsoid containing $\mathcal{E}_1 \cap \Pi_2$. This is the reason why we seek improvements.

The basic idea behind the Fogel-Huang algorithm is to find a family of ellipsoids whose members all contain the intersection of the original ellipsoid and the region between the hyperplanes. Then, the optimal member of the family, in terms of minimising the volume, is selected.

Let $\mathcal{E} = \mathcal{E}(a, Q)$ be the ellipsoid whose (positive definite, symmetric) matrix is Q and centre is a:

$$\theta \in \mathcal{E}(a, Q) \Leftrightarrow (\theta - a)^{\mathrm{T}} Q^{-1} (\theta - a) \leq 1, \tag{3.1}$$

and let $\Pi = \Pi(n, y)$ be the region between two parallel hyperplanes defined by

$$\theta \in \Pi(n, y) \Leftrightarrow (n^{\mathrm{T}}\theta - y)^2 \leq 1 \tag{3.2}$$

where the hyperplanes themselves are defined by

$$\theta \in \mathbb{H}_{\pm}(n, y) \Leftrightarrow y - n^{\mathrm{T}}\theta = \pm 1. \tag{3.3}$$

The family of ellipsoids from which we will select the optimal member is obtained by adding $q(\geq 0)$ times the inequality (3.2) to inequality (3.1):

$$(\theta - a)^{\mathrm{T}} Q^{-1} (\theta - a) + q(n^{\mathrm{T}}\theta - y)^2 \leq 1 + q. \tag{3.4}$$

As the left-hand side of inequality (3.4) is quadratic in θ (and positive definite for nonnegative q) and the right-hand side is constant for constant q, inequality (3.4) defines an ellipsoid $\mathcal{E}(q)$ for each $q \geq 0$. Moreover, any θ satisfying both inequality (3.1) and inequality (3.2) also satisfies inequality (3.4), so $\mathcal{E} \cap \Pi \subset \mathcal{E}(q)$.

Inequality (3.4) can be put in the form

$$(\theta - \bar{a}(q))^{\mathrm{T}} \bar{Q}(q)^{-1} (\theta - \bar{a}(q)) \leq 1, \tag{3.5}$$

by using the matrix inversion lemma, where

$$\bar{Q}(q) = \left(1 + q - \frac{q\nu^2}{1 + gq}\right)\left(I_p - \frac{qQnn^{\mathrm{T}}}{1 + qn^{\mathrm{T}}Qn}\right) Q \tag{3.6}$$

$$\bar{a}(q) = a + \nu q \bar{Q}(q) n, \tag{3.7}$$

where I_p is the identity matrix in the p-dimensional space we assume we are working in, and

$$g = n^{\mathrm{T}} Q n \tag{3.8}$$

$$\nu = y - n^{\mathrm{T}} a, \tag{3.9}$$

and, of course, $\bar{Q}(0) = Q$, $\bar{a}(0) = a$. (g is 4 times the reciprocal of the square of the distance between the hyperplanes, measured in the metric induced by Q and ν is $\sqrt{(n^{\mathrm{T}}n)}$ times the distance between the centre of the ellipsoid and the mid-hyperplane of the hyperplane pair.)

As the volume of an ellipsoid is proportional to the square root of the determinant of its matrix, we are interested in minimising

$$\det \bar{Q}(q) = \left(1 + q - \frac{q\nu^2}{1 + gq}\right)^p \det\left(I_p - \frac{qQnn^{\mathrm{T}}}{1 + qn^{\mathrm{T}}Qn}\right) \det Q$$

$$= \frac{(1 + (1 + g - \nu^2)q + gq^2)^p}{(1 + gq)^{p+1}} \det Q. \tag{3.10}$$

A necessary condition for the quantity in equation (3.10) to be a minimum in the interior of $[0, \infty)$ is that

$$(p-1)g^2q^2 + g(2p-1-g+\nu^2)q + p(1-\nu^2) - g = 0. \tag{3.11}$$

It turns that the quantities

$$a_\pm = \frac{\pm\nu - 1}{\sqrt{g}} \tag{3.12}$$

(the algebraic distances from a to $\mathbb{H}_\pm$ in the metric defined by Q) have a vital rôle in determining whether $\det \bar{Q}(q)$ can be made smaller than $\det Q$.

1. if either $a_+ \geq 1$ or $a_- \geq 1$ then $\mathcal{E} \cap \Pi$ contains at most one point; we assume that this does not happen, so $a_+, a_- < 1$.

2. if $a_+ < -1$ (resp. $a_- < -1$) then $\mathbb{H}_+$ (resp. $\mathbb{H}_-$) does not intersect $\mathcal{E}$. However, in this case, $\mathbb{H}_+$ (resp. $\mathbb{H}_-$) can be replaced by another, parallel, hyperplane tangent to $\mathcal{E}$. This is equivalent to making the transformations $n \to \alpha n$, $y \to \eta$ such that $a_+ \to \tilde{a}_+ = -1$, $a- \to \tilde{a}_- = a_-$ (resp. $a_- \to \tilde{a}_- = -1$, $a+ \to \tilde{a}_+ = a_+$). Under these transformations

$$\begin{array}{ll} g \to \tilde{g} = \frac{4g}{(1+\sqrt{g}+\nu)^2} & \left(\begin{array}{ll} & g \to \tilde{g} = \frac{4g}{(1+\sqrt{g}-\nu)^2} \\ \text{resp.} & \\ & \nu \to \tilde{\nu} = -\frac{1-\sqrt{g}-\nu}{1+\sqrt{g}-\nu} \end{array} \right) \\ \nu \to \tilde{\nu} = \frac{1-\sqrt{g}+\nu}{1+\sqrt{g}+\nu} & \end{array} \tag{3.13}$$

with

$$\tilde{\nu} = 1 - \sqrt{\tilde{g}} \text{ (resp. } \tilde{\nu} = -1 + \sqrt{\tilde{g}}). \tag{3.14}$$

(Actually, the replacement of members of hyperplane pairs which do not intersect the ellipsoid by tangent hyperplanes is necessary to yield a step-optimal algorithm, and this is a modification of the original Fogel-Huang algorithm (Belforte *et al.* [1]). In order to avoid referring to the modified modified algorithm, we will call the Fogel-Huang algorithm with hyperplane shifting the Fogel-Huang algorithm.)

If both $a_+, a_- < -1$, then both hyperplanes could be shifted, and then

$$\begin{array}{rcl} g & \to \tilde{g} & = 1 \\ \nu & \to \tilde{\nu} & = 0 \end{array} . \tag{3.15}$$

However, in this case the hyperplane pair do not exclude any part of the prior ellipsoid, and they are redundant for that reason.

3. $a_+a_- \geq 1/p$. In this case, the smallest value $\det \bar{Q}(q)$ attains on $[0, \infty)$ is $\det Q$ (when $q = 0$).

4. $a_+a_- < 1/p$. In this case, equation (3.11) has a single positive root corresponding to a minimum smaller than $\det Q$.

Figures 3.1 and 3.2 illustrate the various regions in the g-ν plane which are subject to the above transformations and the regions in the $\tilde{g}$-$\tilde{\nu}$ plane to which they are mapped.

We can now summarise the Fogel-Huang algorithm:

Algorithm: 3.1

1. Set the ellipsoid to its initial value $\mathcal{E}_0 = \mathcal{E}(a_0, Q_0)$;

2. for each pair of hyperplanes $\mathbb{H}_i^{\pm}, i = 1, 2, \ldots$, in turn, calculate the indicators a_+ and a_-, using the ellipsoid $\mathcal{E}_{i-1} = \mathcal{E}(a_{i-1}, Q_{i-1})$;

   (a) if $a_+ > 1$ or $a_- > 1$, $\mathcal{E}_i$ would be degenerate, so stop;

   (b) else

      i. if $a_+ < -1$ and $a_- < -1$, the hyperplane pair $\mathbb{H}_i^{\pm}$ is redundant, and we can immediately set $\mathcal{E}_i = \mathcal{E}_{i-1}$;

      ii. else

         A. if $a_+ < -1$, shift $\mathbb{H}_i^+$ to be tangent to $\mathcal{E}_{i-1}$;

         B. if $a_- < -1$, shift $\mathbb{H}_i^-$ to be tangent to $\mathcal{E}_{i-1}$;

         and calculate g and ν for $\mathbb{H}_i^{\pm}$ and $\mathcal{E}_{i-1}$;

         find the positive root q of the quadratic (3.11);

         calculate the determinant of the matrix of the possible new ellipsoid according to equation (3.10)

         A. if this determinant exceeds $\det Q_{i-1}$, set $\mathcal{E}_i = \mathcal{E}_{i-1}$;

         B. otherwise, calculate a_i and Q_i according to equations (3.6) and (3.7) respectively, and set $\mathcal{E}_i = \mathcal{E}(a_i, Q_i)$.

3.1.2 The Minimum Volume

If we actually substitute the positive solution $q_{\min}$ of equation (3.11) in equation (3.10), we find that the ratio of the minimum volume to the original volume is

$$V_r = \sqrt{\frac{\det \bar{Q}\,(q_{\min})}{\det Q}}$$

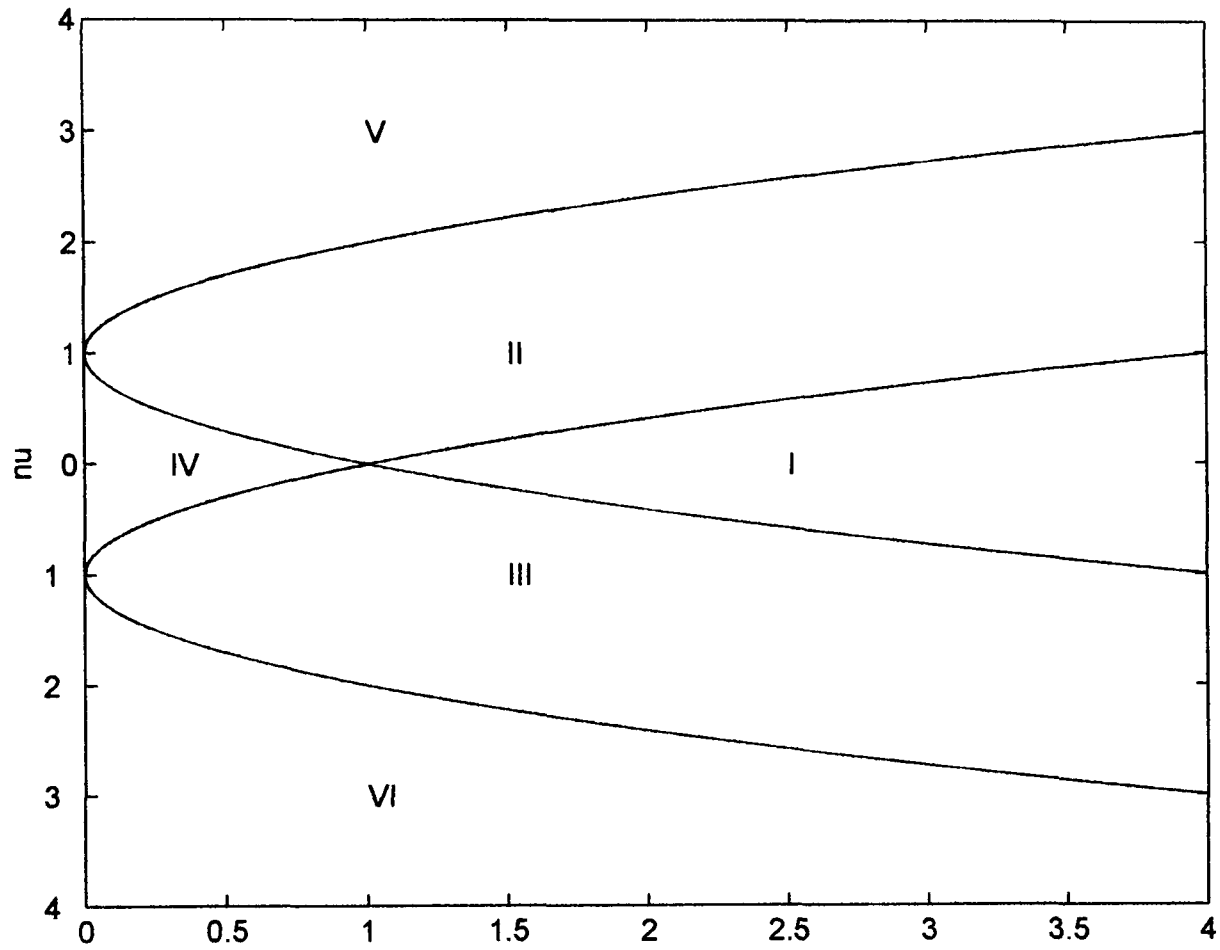


Figure 3.1: the four curves $\nu = \pm 1 \pm' \sqrt{g}$ divide the $g - \nu$ plane into various regions according to whether or not the hyperplanes are shifted to be tangent to the ellipsoid and what transformations they undergo when they are shifted:

Region I corresponds to $a_+, a_- \in (-1, 1)$. Here g, ν remain untransformed;

Region II (resp. III) corresponds to $a_- \leq -1, a_+ \in (-1, 1)$ (resp. $a_+ \leq -1, a_- \in (-1, 1)$). This region is mapped to the upper (resp. lower) boundary of region I;

Region IV corresponds to $a_+, a_- \leq -1$. It is mapped to the point $(1, 0)$;

Region V (resp VI) corresponds to $a_+ \geq 1$ (resp. $a_- \geq 1$). Should not arise.

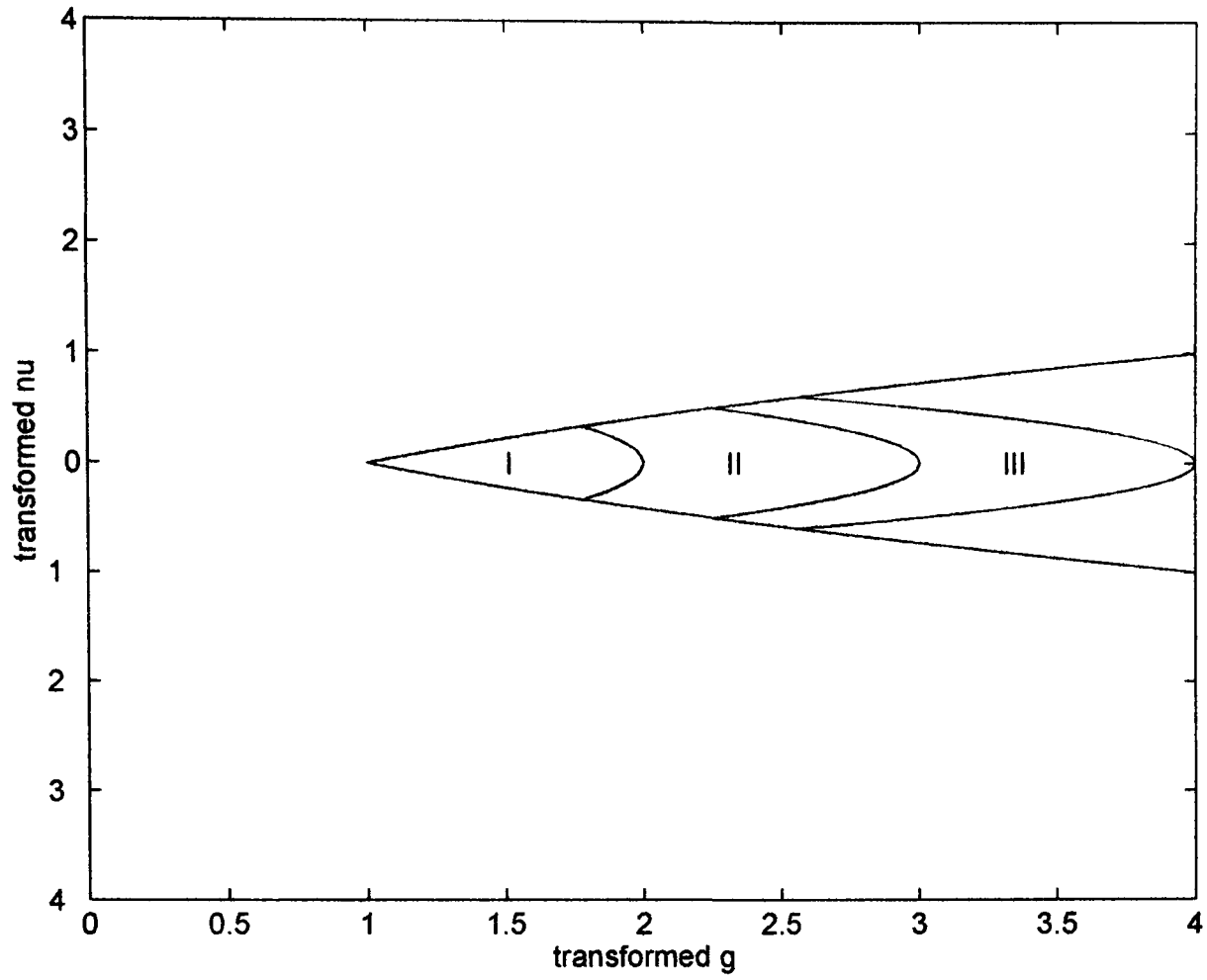


Figure 3.2: The transformed valid values of $(\tilde{g}, \tilde{\nu})$ will be in the wedge shown here. When $|\tilde{\nu}| \leq \sqrt{1-\tilde{g}/p}$, the volume is unchanged. This so in region I when $p \geq 2$, region II when $p \geq 3$, region III when $p \geq 4$, etc.

$$= \begin{cases} \sqrt{\frac{\sqrt{(1-g+\nu^2)^2+4(p^2-1)\nu^2}+(1-g+\nu^2)}{2(p+1)\nu^2}} & \\ \quad \times \left[p \frac{\sqrt{(1-g+\nu^2)^2+4(p^2-1)\nu^2}-p\left(1-g+\nu^2\right)}{g(p^2-1)} \right]^{\frac{p}{2}}, & \nu^2 > 1-\frac{g}{p}, \nu^2 > 0 \\ \sqrt{\frac{p-1}{g-1}} \left(\frac{p(g-1)}{g(p-1)} \right)^{\frac{p}{2}}, & \nu^2 = 0 \geq 1-\frac{g}{p} \\ 1, & \text{otherwise,} \end{cases} \tag{3.16}$$

where we have used Maple©.

Again using Maple©, we find that $\frac{\partial V_{r1}}{\partial \nu} = 0$ (where V_{r1} is the first form of V_r in equation (3.16)) can only happen when $\nu \to 0$, when $\nu^2 = (\sqrt{g}+1)^2$, or when $\nu^2 = 1-g/p$ (if, of course, $g \leq p$). If $g < 1$, $\frac{\partial V_{r1}}{\partial \nu} = 0$ when $\nu^2 = (\sqrt{g}-1)^2$ in addition.

The behaviour of V_{r1} when $\frac{\partial V_{r1}}{\partial \nu} = 0$ is summarised in Table (3.1.2), where these special values of ν are ordered in terms of increasing ν^2.

We see that $V_{r1}(g,\nu)$ is strictly decreasing as a function of ν for each fixed g on $(\sqrt{1-g/p}, 1+\sqrt{g})$ if $g \in (0,p)$, and on $(0, 1+\sqrt{g})$ if $g \geq p$. Consequently, $V_r(g,\nu)$ is monotonically decreasing as a function of ν on $[0, 1+\sqrt{g})$ for all fixed values of g, and, as V_r is even in ν, it is monotonically

g	$\in (0,1]$	$\in (1,p)$	$\in [p,\infty)$
1st $\nu^2 \rightarrow$	0	0	0
$V_{r1} \rightarrow$	∞ $\quad$ $-\infty$ (p even) (p odd)	$\left(\frac{p}{g}\right)^{\frac{p}{2}} \left(\frac{g-1}{p-1}\right)^{\frac{p-1}{2}}$	$\left(\frac{p}{g}\right)^{\frac{p}{2}} \left(\frac{g-1}{p-1}\right)^{\frac{p-1}{2}}$
2nd $\nu^2 =$	$(1-\sqrt{g})^2$	$1-g/p$	$(1+\sqrt{g})^2$
$V_{r1} =$	0	1	0
$\partial^2 V_{r1}/\partial \nu^2 =$	0	negative	0
3rd $\nu^2 =$	$1-g/p$	$(1+\sqrt{g})^2$	—
$V_{r1} =$	1	0	—
$\partial^2 V_{r1}/\partial \nu^2 =$	negative	0	—
4th $\nu^2 =$	$(1+\sqrt{g})^2$	—	—
$V_{r1} =$	0	—	—
$\partial^2 V_{r1}/\partial \nu^2 =$	0	—	—

Table 3.1: Special Values of ν^2

increasing on $(-1-\sqrt{g}, 0]$. But $a_-, a_+ < 1 \Rightarrow \nu \in (-1-\sqrt{g}, 1+\sqrt{g})$, so, in the region of interest, V_r is monotonically decreasing as the magnitude of ν increases.

It is desirable to have a measure of the size of a bounded convex body, representing a feasible region in parameter space, which more directly reflects the uncertainty of the individual parameters, no matter what the dimension of the parameter space.

Definition 3.1: *The* characteristic length *of a bounded convex body in* $\mathbb{R}^p$ *is the pth root of its volume.* □

Clearly, for an ellipsoid, the characteristic length is proportional to the geometric mean of its semi-axes, and for an axis-aligned ellipsoid this in turn is proportional to the geometric mean of the individual parameter uncertainties.

Casting equation (3.16) in terms of the ratio of the characteristic lengths of the ellipsoids, $C_r = V_r^{1/p}$, we have

$$C_r = \left(\frac{\det \bar{Q}\left(q_{\min}\right)}{\det Q}\right)^{1/2p}$$

$$= \begin{cases} \sqrt{p\frac{\sqrt{(1-g+\nu^2)^2+4(p^2-1)\nu^2}-p(1-g+\nu^2)}{g(p^2-1)}} & \\ \quad \times \left[\frac{\sqrt{(1-g+\nu^2)^2+4(p^2-1)\nu^2}+\left(1-g+\nu^2\right)}{2(p+1)\nu^2}\right]^{1/2p}, & \nu^2 > 1-\frac{g}{p}, 0 \\ \sqrt{\frac{p(g-1)}{g(p-1)}}\left(\frac{p-1}{g-1}\right)^{1/2p} & \nu^2 = 0 \geq 1-\frac{g}{p} \\ 1, & \text{otherwise.} \end{cases} \tag{3.17}$$

3.2 Data for a Monte-Carlo Test of the Performance of the Fogel-Huang Algorithm

We use MATLAB©-generated data[1] for the model

$$y_k = \hat{\theta}^1 y_{k-1} + \cdots + \hat{\theta}^{p-1} y_{k-p+1} + \hat{\theta}^p u_k + v_k \tag{3.18}$$

where $\hat{\theta} = (1, 1, \ldots, 1)^{\mathrm{T}}/\sqrt{p}$ is the "true" value of the parameter vector, $y_{2-p}, \ldots, y_0$ are randomly selected in $[-1, 1]$ (with either a uniform distribution (with standard deviation $\sigma_u = \frac{1}{\sqrt{3}}$) or truncated normal distribution with standard deviation[2] $\sigma_t = \frac{1}{2}\sigma_u$ or $\frac{1}{4}\sigma_u$), and v_k is selected

[1]When MATLAB©'s random number generators were checked for autocorrelation, the following results were obtained: for 1000 sequences of 10,000 zero-mean uniformly distributed random numbers the autocorrelation at lag 1 exceeded $2\sqrt{10,000}$ times the autocorrelation at lag 0 in 24 cases and exceeded $3\sqrt{10,000}$ times the autocorrelation at lag 0 in 3 cases. The autocorrelation at lag 2 exceeded the values relative to the 0-lag autocorrelation in 20 cases and 1 case respectively. Similar results were found for for 1000 sequences of 10,000 zero-mean normally distributed random numbers with standard deviation 1. The expected results for truly uncorrelated sequences (see, e.g., Norton [15]) would be 9 or 10 cases exceeding $2\sqrt{10,000}$ times the autocorrelation at lag 0 and 0 or 1 case exceeding $3\sqrt{10,000}$ times the autocorrelation at lag 0, both when the autocorrelation at lag 1 and that at lag 2 are evaluated, so the MATLAB©-generated random numbers are reasonably uncorrelated.

[2]That is, if σ is the standard deviation of the zero-mean normal distribution before truncation, then

$$\sigma_t^2 = \sigma^2 - \sqrt{\frac{2}{\pi}}\frac{e^{-\frac{1}{2}\sigma^2}}{\operatorname{erf}\left(\frac{1}{\sigma\sqrt{2}}\right)}\sigma.$$

in the same way as these y's. The input u_k is 1 for all $k > 0$.

Thus, the kth regressor is given by

$$n_k = (n_k^1, \ldots, n_k^p)^{\mathrm{T}} = (y_{k-1}, \ldots, y_{k-p+1}, 1)^{\mathrm{T}}. \tag{3.19}$$

We use equation (3.18) for $k = 1, \ldots, 12$ so there are 12 hyperplane pairs for each case, and we run the computer programs implementing the Fogel-Huang and other algorithms for 100 different realisations of each noise distribution (and different $y_{2-p}, \ldots, y_0$) for each of $p = 2, \ldots, 5$. In each case, the initial prior set is $\{\theta : \theta^{\mathrm{T}}\theta \leq 20^2\}$, this set being chosen to guarantee the inclusion of both our $\hat{\theta}$, and all the vertices of the polytope enclosed by the set of hyperplane pairs. This is done to avoid the complication of part of the polytope being excluded by the initial prior set and is achieved by simply calculating the vertices of the polytope for each data set and choosing the radius of the initial prior set large enough to contain them all.

It will be noted that the bound specification $-1 \leq y_k - n^{\mathrm{T}}\hat{\theta} \leq 1$ is more general than it appears at first sight. For, if $e_k^m \leq y_k - \hat{\theta}^{\mathrm{T}} n_k \leq e_k^M$, then $-1 \leq \bar{y}_k - \hat{\theta}^{\mathrm{T}} \bar{n}_k \leq 1$, where the transformed data $\bar{y}_k = (2y_k - e_k^m - e_k^M)/(e_k^M - e_k^m)$, $\bar{n}_k = 2n_k/(e_k^M - e_k^m)$ (see Pronzato and Walter [22]).

In addition, this model is dynamic according to the classification of Norton and Veres[24], in that the same variables appear both as outputs and (more than once, if $p > 2$) in the regressors. This will produce some complications below.

3.2.1 Expected Behaviour of F-H Characteristic Lengths with Uniform v_k and $y_{2-p} \ldots y_0$.

We generalise to the plant model

$$\begin{aligned} y_k &= -\sum_{i=1}^{r-1} a^i y_{k-r+i} + \sum_{i=1}^{s} b^i u_{k-s+i} + v_k \\ &= \hat{\theta}^{\mathrm{T}} n_k + v_k \end{aligned}$$

where

- y_k is the kth output,
- u_j is the jth input
- v_k is noise, uniformly distributed in $[-1, 1]$,

- $\hat{\theta} = (-a^1, \ldots, -a^{r-1}, b^1, \ldots, b^s) \in \mathbb{R}^p$ is a vector of parameters ($p = r + s - 1$),
- and $n_k = (y_{k-r+1}, \ldots, y_{k-1}, u_{k-s+1}, \ldots, u_k)$ is the kth regressor

(step number is indexed by subscripts, components of vectors by superscripts).

Assume that we have the Fogel-Huang ellipsoid $\mathcal{E}(\theta_k, Q_k) = \{\theta : (\theta - \theta_k)^{\mathrm{T}} Q_k^{-1} (\theta - \theta_k) \leq 1\} \ni \hat{\theta}$ and further assume that the actual parameter vector $\hat{\theta}$ is equally likely to be any point of $\mathcal{E}(\theta_{k-1}, Q_{k-1})$.

Let us make a shift of origin $k \to 1$, so that we are considering the effect of the data $(y_1, n_1) = (y_1, (y_{2-r}, \ldots, y_0, u_{2-s}, \ldots, u_1))$ on the ellipsoid $\mathcal{E}(\theta_0, Q_0)$.

Then, the *a priori* probability density function $p_\theta(\theta)$ of θ is given by

$$p_\theta(\theta) = \begin{cases} U_p^{-1} \left(\det(Q_0)\right)^{-\frac{1}{2}} & \text{if } \theta \in \mathcal{E}(\theta_0, Q_0) \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

where

$$\begin{aligned} U_p &= \mathrm{B}\left(\frac{1}{2}, 1\right) \mathrm{B}\left(\frac{1}{2}, \frac{3}{2}\right) \cdots \mathrm{B}\left(\frac{1}{2}, \frac{(p+1)}{2}\right) \\ &= \frac{\Gamma(\frac{1}{2})\Gamma(1)}{\Gamma(\frac{3}{2})} \frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(2)} \cdots \frac{\Gamma(\frac{1}{2})\Gamma(\frac{p+1}{2})}{\Gamma(\frac{1}{2}p+1)} \\ &= \frac{\Gamma(\frac{1}{2})^p}{\Gamma(\frac{1}{2}p+1)} = \frac{\pi^{p/2}}{\Gamma(\frac{1}{2}p+1)} \end{aligned}$$

is the volume of the unit sphere in p dimensions.

Thus

$$\begin{aligned} \Pr(\theta^{\mathrm{T}} n_1 \leq z) &= \frac{1}{U_p (\det Q_0)^{\frac{1}{2}}} \int \cdots \int_{(\theta-\theta_0)^{\mathrm{T}} Q_0^{-1} (\theta-\theta_0) \leq 1, \theta^{\mathrm{T}} n_1 \leq z} \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^p \\ &= \frac{1}{U_p} \int \cdots \int_{\theta^{\mathrm{T}}\theta \leq 1, \theta^{\mathrm{T}} Q_0^{\frac{1}{2}} n_1 \leq z - \theta_0{}^{\mathrm{T}} n_1} \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^p \\ &\qquad \text{(under the transformation } \theta \to \theta_0 + Q_0^{\frac{1}{2}}\theta) \\ &= \frac{1}{U_p} \int \cdots \int_{\theta^{\mathrm{T}}\theta \leq 1, \theta^p \leq (z - \theta_0{}^{\mathrm{T}} n_1)/\sqrt{n_1{}^{\mathrm{T}} Q_0 n_1}} \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^p \\ &\qquad \text{(under the orthogonal transformation which makes } Q_0^{\frac{1}{2}} n_1 \\ &\qquad \text{parallel to the } p\text{th coordinate axis)} \\ &= \frac{1}{U_p} \int_{-1}^{(z - \theta_0{}^{\mathrm{T}} n_1)/\sqrt{g_1}} \int_{-\sqrt{1-(\theta^p)^2}}^{\sqrt{1-(\theta^p)^2}} \cdots \int_{-\sqrt{1-\sum_{j=2}^p (\theta^j)^2}}^{\sqrt{1-\sum_{j=2}^p (\theta^j)^2}} \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^{p-1} \mathrm{d}\theta^p \end{aligned}$$

or

$$\Pr(\theta^{\mathrm{T}} n_1 \leq z) = \frac{1}{\mathrm{B}(\frac{1}{2}, \frac{p+1}{2})} \int_{-1}^{(z-\theta_0{}^{\mathrm{T}} n_1)/\sqrt{g_1}} (1-(\theta^p)^2)^{(p-1)/2} \mathrm{d}\theta^p \tag{3.21}$$

where $Q_0^{\frac{1}{2}}$ is any square root of Q_0 (which, of course, exists as Q_0 is positive definite). (We have attached a subscript to g_1, the "g" parameter of the Fogel-Huang algorithm, to distinguish it from the same parameter for the next step.)

Differentiating, we find

$$p_{(\theta-\theta_0)^{\mathrm{T}} n_1}(z) = \begin{cases} \frac{1}{\sqrt{g_1}\mathrm{B}(\frac{1}{2}, \frac{p+1}{2})} \left[1 - \frac{z^2}{g_1}\right]^{(p-1)/2}, & z \in [-\sqrt{g_1}, \sqrt{g_1}] \\ 0, & \text{otherwise} \end{cases} \tag{3.22}$$

We wish to use this expression to find the distribution of $\nu_1 = (\theta - \theta_0)^{\mathrm{T}} n_1 + v_1$ (where ν_1 has a subscript for the same reason that g_1 has). There are two cases:

$g_1 \leq 1$ Here

$$\Pr(\nu_1 \leq w) = \begin{cases} 0, & \text{if } w \leq -1 - \sqrt{g_1} \\ \frac{1}{2\sqrt{g_1}\mathrm{B}\left(\frac{1}{2}, \frac{1}{2}(p+1)\right)} \int_{-\sqrt{g_1}}^{w+1} \int_{-1}^{w-z} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}v\, \mathrm{d}z, & \text{if } w \in [-1-\sqrt{g_1}, -1+\sqrt{g_1}] \\ \frac{1}{2\sqrt{g_1}\mathrm{B}\left(\frac{1}{2}, \frac{1}{2}(p+1)\right)} \int_{-\sqrt{g_1}}^{\sqrt{g_1}} \int_{-1}^{w-z} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}v\, \mathrm{d}z, & \text{if } w \in [-1+\sqrt{g_1}, 1-\sqrt{g_1}] \\ 1 - \frac{1}{2\sqrt{g_1}\mathrm{B}\left(\frac{1}{2}, \frac{p+1}{2}\right)} \int_{w-1}^{\sqrt{g_1}} \int_{w-z}^{1} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}v\, \mathrm{d}z, & \text{if } w \in [1-\sqrt{g_1}, 1+\sqrt{g_1}] \\ 1, & \text{if } w \geq 1 + \sqrt{g_1} \end{cases} \tag{3.23}$$

so, by differentiation,

$$p_{\nu_1}(\nu_1) = \frac{\mathrm{d}}{\mathrm{d}w}\,\Pr(\nu_1 \le w)\Big|_{w=\nu_1} = \begin{cases} 0, & \text{if } \nu_1 \notin [-1-\sqrt{g_1}, 1+\sqrt{g_1}], \\ \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \int_{-\sqrt{g_1}}^{\nu_1+1} \left[1-\frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}z, & \\ \qquad \text{if } \nu_1 \in [-1-\sqrt{g_1}, -1+\sqrt{g_1}] & \\ \frac{1}{2}, & \text{if } \nu_1 \in [-1+\sqrt{g_1}, 1-\sqrt{g_1}], \\ \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \int_{\nu_1-1}^{\sqrt{g_1}} \left[1-\frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}z, & \\ \qquad \text{if } \nu_1 \in [1-\sqrt{g_1}, 1+\sqrt{g_1}], & \end{cases}$$

or, on making the substitution $z = \sqrt{g_1}(1-2t)$ and noting that $\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right) = 2^p \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)$,

$$p_{\nu_1}(\nu_1) = \begin{cases} 0, & \text{if } \nu_1 \notin [-1-\sqrt{g_1}, 1+\sqrt{g_1}], \\ \frac{1}{2}, & \text{if } \nu_1 \in [-1+\sqrt{g_1}, 1-\sqrt{g_1}], \\ \frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1+\frac{1-|\nu_1|}{\sqrt{g_1}}\right)\right)}{2\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}, & \text{otherwise}, \end{cases} \tag{3.24}$$

where

$$\mathrm{B}(r,s,z) = \int_0^z t^{r-1}(1-t)^{s-1}\mathrm{d}t, r, s > 0, z \in [0,1] \tag{3.25}$$

is the incomplete beta function.

$g_1 > 1$ Here

$$\Pr(\nu_1 \le w) = \begin{cases} 0, & \text{if } w \le -1 - \sqrt{g_1} \\ \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \int_{-\sqrt{g_1}}^{w+1} \int_{-1}^{w-z} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}v\,\mathrm{d}z, & \\ \qquad \text{if } w \in [-1 - \sqrt{g_1}, 1 - \sqrt{g_1}] & \\ \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \left\{ \int_{-\sqrt{g_1}}^{w-1} \int_{-1}^{1} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}v\,\mathrm{d}z + \right. & \\ \qquad \left. \int_{w-1}^{w+1} \int_{-1}^{w-z} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}v\,\mathrm{d}z \right\}, & \\ \qquad \text{if } w \in [1 - \sqrt{g_1}, -1 + \sqrt{g_1}] & \\ 1 - \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \int_{w-1}^{\sqrt{g_1}} \int_{w-z}^{1} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}v\,\mathrm{d}z, & \\ \qquad \text{if } w \in [-1 + \sqrt{g_1}, 1 + \sqrt{g_1}] & \\ 1, & \text{if } w \ge 1 + \sqrt{g_1} \end{cases} \tag{3.26}$$

so, again by differentiation,

$$p_{\nu_1}(\nu_1) = \begin{cases} 0, & \text{if } \nu_1 \notin [-1 - \sqrt{g_1}, 1 + \sqrt{g_1}]; \\ \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \int_{-\sqrt{g_1}}^{\nu_1+1} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}z, & \\ \qquad \text{if } \nu_1 \in [-1 - \sqrt{g_1}, 1 - \sqrt{g_1}]; & \\ \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \int_{\nu_1-1}^{\nu_1+1} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}z, & \\ \qquad \text{if } \nu_1 \in [1 - \sqrt{g_1}, -1 + \sqrt{g_1}]; & \\ \frac{1}{2\sqrt{g_1}\,\mathrm{B}\left(\frac{1}{2},\frac{p+1}{2}\right)} \int_{\nu_1-1}^{\sqrt{g_1}} \left[1 - \frac{z^2}{g_1}\right]^{\frac{p-1}{2}} \mathrm{d}z, & \\ \qquad \text{if } \nu_1 \in [-1 + \sqrt{g_1}, 1 + \sqrt{g_1}], & \end{cases}$$

and, putting $z = \sqrt{g_1}(1 - 2t)$ once more

$$p_{\nu_1}(\nu_1) = \begin{cases} 0, & \text{if } \nu_1 \notin [-1 - \sqrt{g_1}, 1 + \sqrt{g_1}]; \\ \frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\nu_1-1}{\sqrt{g_1}}\right)\right) - \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\nu_1+1}{\sqrt{g_1}}\right)\right)}{2\,\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}, & \\ \qquad \text{if } \nu_1 \in [1 - \sqrt{g_1}, -1 + \sqrt{g_1}] & \\ \frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1+\frac{1-|\nu_1|}{\sqrt{g_1}}\right)\right)}{2\,\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}, & \text{otherwise.} \end{cases} \tag{3.27}$$

(Since $y_1 = \theta^{\mathrm{T}} n_1 + v_1 = \nu_1 + \theta_0^{\mathrm{T}} n_1$, the probability density function for y_1, $p_{y_1}(y_1) = p_{\nu_1}(y_1 - \theta_0^{\mathrm{T}} n_1)$.)

We now wish to find the probability density functions of $\tilde{g}_1$ and $\tilde{\nu}_1$, that is, of g_1 and ν_1 as transformed according to (3.13) and (3.15):

Case (1). g_1 and ν_1 are not transformed if $\nu_1 \in [-\sqrt{g_1} + 1, \sqrt{g_1} - 1]$, which requires $g_1 > 1$;

Case (2). g_1 and ν_1 are transformed according to equations (3.13) if $\nu_1 \in [-1 - \sqrt{g_1}, \min[-1 + \sqrt{g_1}, 1 - \sqrt{g_1}])$;

Case (3). g_1 and ν_1 are also transformed according to equations (3.13) if $\nu_1 \in (\max[-1 + \sqrt{g_1}, 1 - \sqrt{g_1}], 1 + \sqrt{g_1}]$;

Case (4). g_1 and ν_1 are transformed according to equations (3.15) if $\nu_1 \in [-1 + \sqrt{g_1}, 1 - \sqrt{g_1}]$ (which requires $g_1 \leq 1$).

The probability of Case (4) is given by:

$$\Pr(\text{Case (4)}) = \begin{cases} \int_{-1+\sqrt{g_1}}^{1-\sqrt{g_1}} p_{\nu_1}(\nu) \mathrm{d}\nu = \int_{-1+\sqrt{g_1}}^{1-\sqrt{g_1}} \frac{1}{2} \mathrm{d}\nu = 1 - \sqrt{g_1}, & \text{if } g_1 \in (0, 1] \\ 0, & \text{otherwise.} \end{cases} \tag{3.28}$$

The probabilities of the other cases are not needed for the calculation of $p_{\tilde{g}_1 \tilde{\nu}_1}(\tilde{g}_1, \tilde{\nu}_1)$, but we will calculate them anyway.

The probability of Case (2) (which corresponds to the hyperplane $\mathbb{H}^+$ of the first pair failing to cut the ellipsoid while $\mathbb{H}^-$ cuts it) is given by:

$g_1 \leq 1$:

$$\begin{aligned} &\Pr(\text{Case (2)}) \\ &= \frac{1}{2\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} \int_{-\sqrt{g_1}-1}^{\sqrt{g_1}-1} \mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, \frac{1}{2}\left(1 + \frac{1+\nu}{\sqrt{g_1}}\right)\right) \mathrm{d}\nu \end{aligned}$$

or

$$\Pr(\text{Case (2)}) = \frac{\sqrt{g_1}}{2}, \tag{3.29}$$

by the substitution $\nu = 2\sqrt{g_1} x - \sqrt{g_1} - 1$, equation (3.24) and the fact that

$$\begin{aligned} \int_0^x \mathrm{B}(r, s, t) \mathrm{d}t &= x\mathrm{B}(r, s, x) - \mathrm{B}(r+1, s, x) \\ &= \left(x - \frac{r}{r+s}\right) \mathrm{B}(r, s, x) + \frac{x^r (1-x)^s}{r+s}, \end{aligned} \tag{3.30}$$

so $\int_0^1 \mathrm{B}(r,s,t)\mathrm{d}t = \frac{s}{r+s}\mathrm{B}(r,s)$.

Equation (3.30) is derived by using integration by parts twice, once to generate the recurrence relation

$$\mathrm{B}(r,s,x) = \frac{r-1}{r+s-1}\mathrm{B}(r-1,s,x) - \frac{x^{r-1}(1-x)^s}{r+s-1} \tag{3.31}$$

for the beta function, and once directly.

$g_1 \geq 1$:

$$\begin{aligned}
&\Pr\,(\text{Case }(2)) \\
&\quad = \frac{1}{2\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)} \int_{-\sqrt{g_1}-1}^{-\sqrt{g_1}+1} \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1+\frac{1+\nu}{\sqrt{g_1}}\right)\right)\mathrm{d}\nu \\
&\quad = \frac{(2-\sqrt{g_1})\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},g_1^{-\frac{1}{2}}\right)}{2\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)} + \frac{\sqrt{g_1}}{(p+1)\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\left(\frac{1}{\sqrt{g_1}}-\frac{1}{g_1}\right)^{(p+1)/2},
\end{aligned} \tag{3.32}$$

by the definition of the incomplete beta function.

By the definition of the complete beta function, the right-hand side of equation (3.32) tends to $\frac{1}{2}$ as $g \to 1$, as we would like it to, as when the hyperplanes have separation 2 in the metric in which the ellipsoid is the unit sphere, the probability of a particular member of the hyperplane pair intersecting the ellipsoid is $\frac{1}{2}$ and the probability of both hyperplanes intersecting is 0.

Case (3) (which corresponds to the hyperplane $\mathbb{H}^-$ of the first pair failing to cut the ellipsoid while $\mathbb{H}^+$ cuts it) is dual to Case (2) so Pr(Case (3)) = Pr(Case (2)).

Case (1) (corresponding to both hyperplanes from the first pair cutting the ellipsoid) consequently has a probability given by Pr(Case (1)) = 1 − Pr(Case (4)) − 2 Pr(Case (2)):

$$\Pr(\text{Case }(1)) = \begin{cases} 0, & \text{if } g_1 \in (0,1] \\ 1 - \frac{(2-\sqrt{g_1})\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},g_1^{-\frac{1}{2}}\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)} & \\ \quad - \frac{2\sqrt{g_1}}{(p+1)\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\left(\frac{1}{\sqrt{g_1}}-\frac{1}{g_1}\right)^{(p+1)/2} & \text{if } g_1 > 1 \end{cases} \tag{3.33}$$

These probabilities are illustrated in Figures (3.3) and (3.4). (We note that when $g_1 = 1$, so that the hyperplanes are separated by 2 in the metric in which the ellipsoid is a unit sphere, Pr(Case (2) or Case (3)) = 1, i.e., the *probability* that exactly one hyperplane intersects the ellipsoid is 1 — although, geometrically, both hyperplanes might touch the ellipsoid).

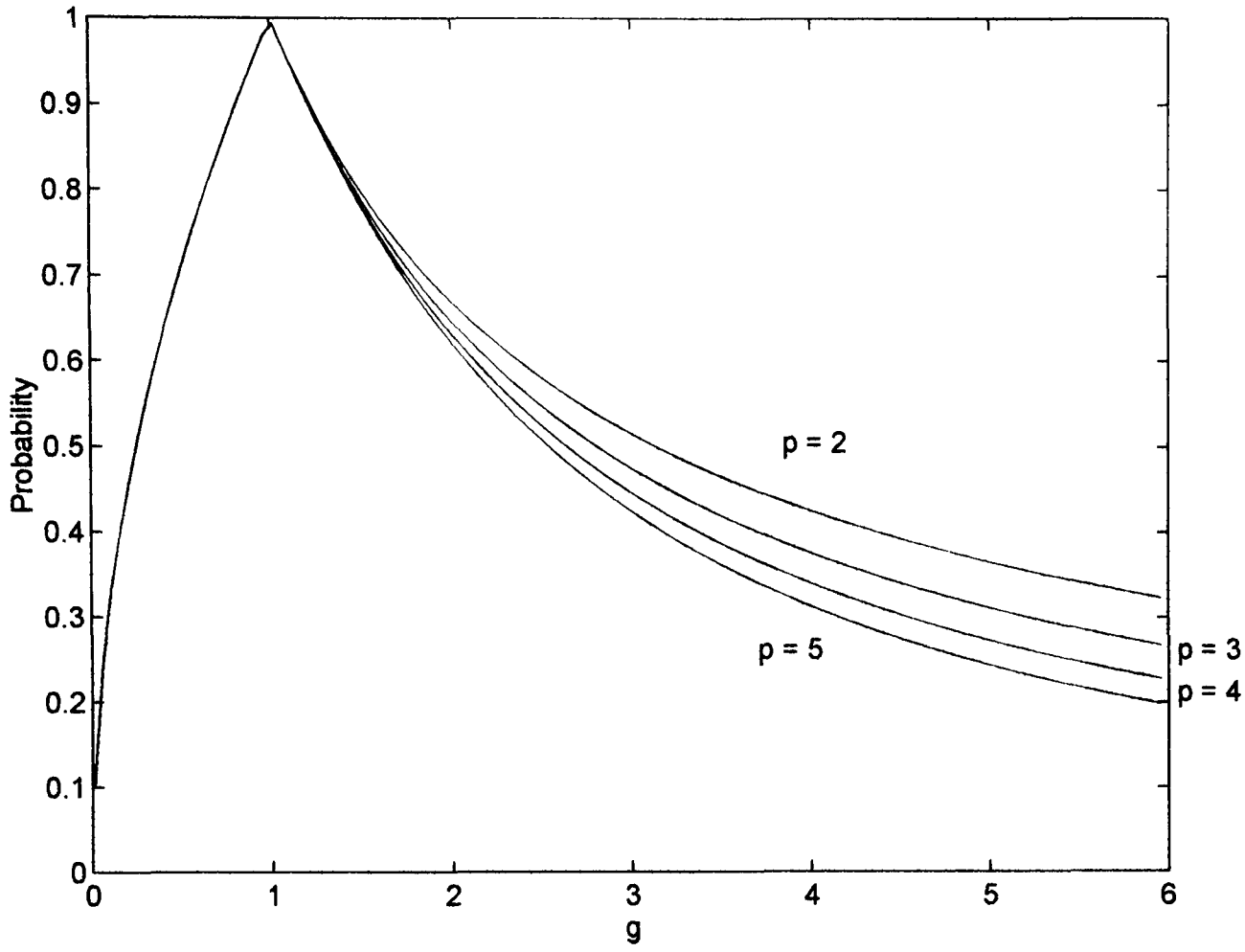


Figure 3.3: Probability that exactly one hyperplane intersects the ellipsoid.

We now return to the task of finding $p_{\tilde{g}_1\tilde{\nu}_1}(\tilde{g}_1, \tilde{\nu}_1)$.

$g_1 \leq 1$: for fixed g_1, equations (3.13), (3.14) and (3.15) can be considered functions of the single variable ν_1, and then, with Case (4), we have

$$
\begin{aligned}
& p_{\tilde{g}_1\tilde{\nu}_1}(\tilde{g}_1, \tilde{\nu}_1) \\
&= \Pr(\nu_1 \in [-1+\sqrt{g_1}, 1-\sqrt{g_1}])\delta(\tilde{g}_1 - 1)\delta(\tilde{\nu}_1) \\
&\quad + \left(\left| \frac{\mathrm{d}\tilde{g}_1}{\mathrm{d}\nu_1} \right|^{-1}_{\tilde{\nu}_1 > 0} p_{\nu_1}(\nu_1)\delta\left(\tilde{\nu}_1 - 1 + \sqrt{\tilde{g}_1}\right) \right) + \\
&\qquad \left(\left| \frac{\mathrm{d}\tilde{g}_1}{\mathrm{d}\nu_1} \right|^{-1}_{\tilde{\nu}_1 < 0} p_{\nu_1}(\nu_1)\delta\left(\tilde{\nu}_1 + 1 - \sqrt{\tilde{g}_1}\right) \right) \\
&= (1-\sqrt{g_1})\delta(\tilde{g}_1 - 1)\delta(\tilde{\nu}_1) \\
&\quad + \left(\frac{g_1}{\tilde{g}_1^3}\right)^{\frac{1}{2}} \left\{ p_{\nu_1}\left(2\left[\frac{g_1}{\tilde{g}_1}\right]^{\frac{1}{2}} - 1 - \sqrt{g_1} \right) \delta\left(\tilde{\nu}_1 + 1 - \sqrt{\tilde{g}_1}\right) \right. \\
&\qquad \left. + p_{\nu_1}\left(-2\left[\frac{g_1}{\tilde{g}_1}\right]^{\frac{1}{2}} + 1 + \sqrt{g_1} \right) \delta\left(\tilde{\nu}_1 - 1 + \sqrt{\tilde{g}_1}\right) \right\},
\end{aligned}
$$

(where the appropriate branches of the equations inverting equations (3.13) have been used in each occurrence of $|\mathrm{d}\tilde{g}_1/\mathrm{d}\nu_1|^{-1} p_{\nu_1}(\nu_1)$) and then, using equation (3.24)

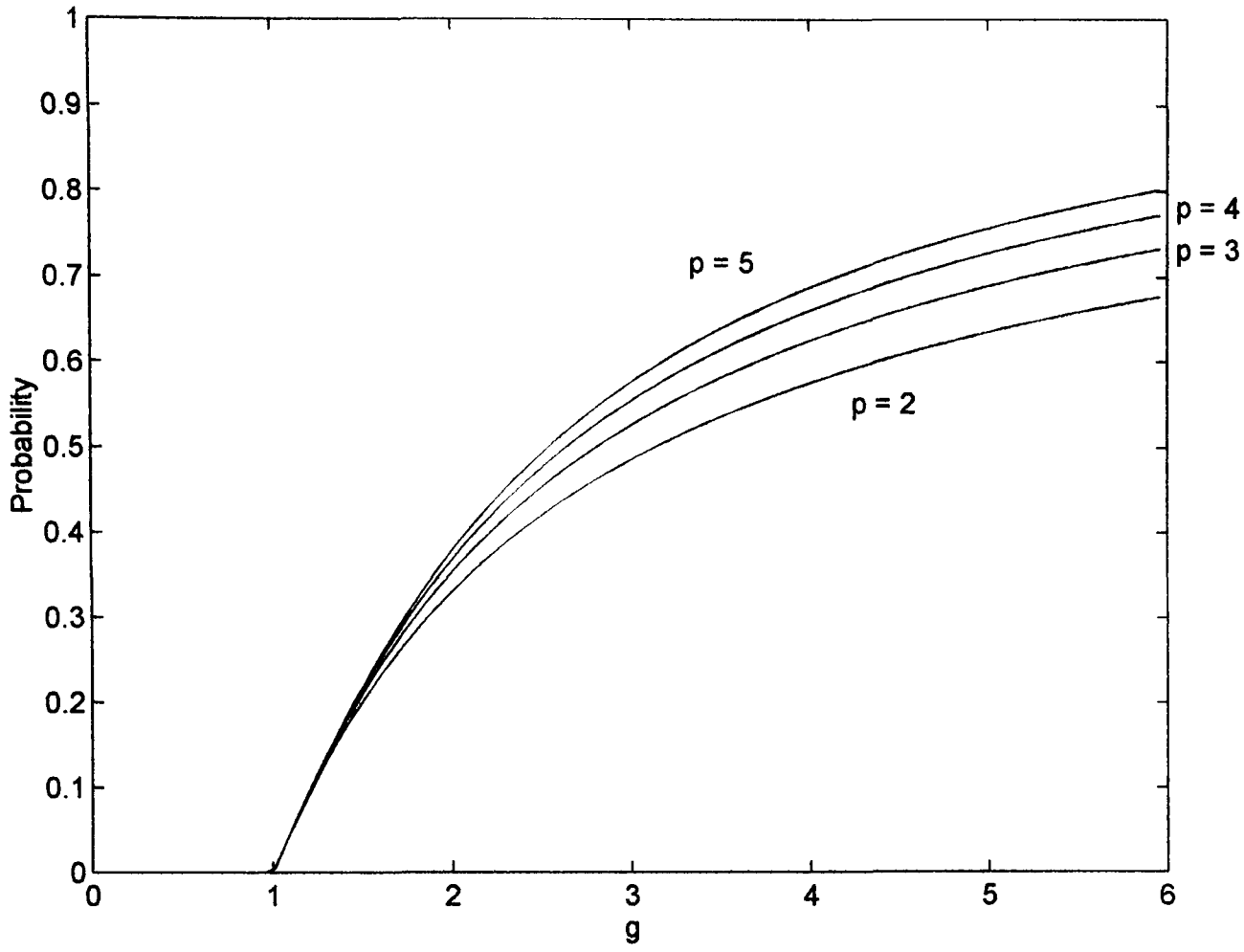


Figure 3.4: Probability that two hyperplanes intersect the ellipsoid.

$$
\begin{aligned}
& p_{\tilde{g}_1 \tilde{\nu}_1}(\tilde{g}_1, \tilde{\nu}_1) \\
&= (1 - \sqrt{g_1})\delta(\tilde{g}_1 - 1)\delta(\tilde{\nu}_1) \\
&\quad + \frac{1}{2}\left(\frac{g_1}{\tilde{g}_1^3}\right)^{\frac{1}{2}} \frac{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, \tilde{g}_1^{-\frac{1}{2}}\right)}{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} \left[\delta\left(\tilde{\nu}_1 + 1 - \sqrt{\tilde{g}_1}\right) + \delta\left(\tilde{\nu}_1 - 1 + \sqrt{\tilde{g}_1}\right)\right]
\end{aligned}
\tag{3.34}
$$

where δ is the Dirac delta "function".

$g_1 > 1$: Again using equations (3.13), (3.14), and (3.15), we have:

$$p_{\tilde{g}_1\tilde{\nu}_1}(\tilde{g}_1,\tilde{\nu}_1) = \begin{cases} 0, & \tilde{g}_1 < g_1; \\ p_{\nu_1}(\tilde{\nu}_1)\delta(g_1-\tilde{g}_1), & |\tilde{\nu}_1| < |\sqrt{\tilde{g}_1}-1|; \\ \left(\dfrac{g_1}{\tilde{g}_1^3}\right)^{\frac{1}{2}} \left\{ p_{\nu_1}\left(2\left[\dfrac{g_1}{\tilde{g}_1}\right]^{\frac{1}{2}} - 1 - \sqrt{g_1}\right)\delta(\tilde{\nu}_1+1-\sqrt{\tilde{g}_1}) \right. & \\ \left. \quad + p_{\nu_1}\left(-2\left[\dfrac{g_1}{\tilde{g}_1}\right]^{\frac{1}{2}} + 1 + \sqrt{g_1}\right)\delta(\tilde{\nu}_1-1,+\sqrt{\tilde{g}_1}) \right\} & \\ & \text{otherwise,} \end{cases}$$

and, then, using equations (3.28), and (3.27),

$$p_{\tilde{g}_1\tilde{\nu}_1}(\tilde{g}_1,\tilde{\nu}_1)$$

$$= \begin{cases} 0, & \tilde{g}_1 < g_1; \\ \dfrac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\tilde{\nu}_1-1}{\sqrt{g_1}}\right)\right) - \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\tilde{\nu}_1+1}{\sqrt{g_1}}\right)\right)}{2\,\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\delta(\tilde{g}_1-g_1), & \\ & \text{if } |\tilde{\nu}_1| < \sqrt{\tilde{g}_1}-1; \\ \dfrac{1}{2}\left(\dfrac{g_1}{\tilde{g}_1^3}\right)^{\frac{1}{2}} \dfrac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\tilde{g}_1^{-\frac{1}{2}}\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}[\delta(\tilde{\nu}_1+1-\sqrt{\tilde{g}_1})+\delta(\tilde{\nu}_1-1+\sqrt{\tilde{g}_1})], & \\ & \text{otherwise.} \end{cases} \tag{3.35}$$

The probability that the algorithm leaves the characteristic length unchanged is

$$\int_{1-}^{p}\int_{-\sqrt{1-g/p}}^{\sqrt{1-g/p}} p_{\tilde{g}_1\tilde{\nu}_1}(g,\nu)\mathrm{d}\nu\mathrm{d}g, \tag{3.36}$$

(where the limiting process implied by the integration bound "1−" is necessary to include the points in the $g_1\nu_1$ plane which are mapped to $(1,0)$ in the $\tilde{g}_1\tilde{\nu}_1$ plane) by equation (3.16)). We again have to treat various ranges of g_1 separately.

$g_1 \in (0,1]$: Here,

$$\Pr(C_r=1) \;=\; 1-\sqrt{g_1}+\sqrt{g_1}\int_1^{4p^2/(p+1)^2} g^{-\frac{3}{2}}\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},g^{-\frac{1}{2}}\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}g$$

$$
\begin{aligned}
&(\text{as } 1-\sqrt{g} < \sqrt{1-g/p} \text{ requires } g < 4p^2/(p+1)^2) \\
&= 1-\sqrt{g_1}+2\sqrt{g_1}\int_{(p+1)/2p}^{1} \frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},x\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}x \\
&= 1-\sqrt{g_1}\left[\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{p+1}{2p}\right)}{p\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}+\frac{(p+1)^{\frac{1}{2}(p-1)}(p-1)^{\frac{p+1}{2}}}{2^p p^{p+1}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\right],
\end{aligned}
\tag{3.37}
$$

from equations (3.36), (3.34), and the expression from equation (3.30) for $\int_0^x \mathrm{B}(r,s,z)\mathrm{d}z$.

$g_1 \in [1, 4p^2/(p+1)^2]$:

The corresponding probability here is

$$
\begin{aligned}
\Pr(C_r=1) &= \int_0^{\sqrt{g_1}-1} \frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\nu-1}{\sqrt{g_1}}\right)\right)-\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\nu+1}{\sqrt{g_1}}\right)\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}\nu \\
&\quad + \sqrt{g_1}\int_{g_1}^{4p^2/(p+1)^2} g^{-\frac{3}{2}}\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},g^{-\frac{1}{2}}\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}g \\
&= 2\sqrt{g_1}\left(\int_{g_1^{-\frac{1}{2}}}^{\frac{1}{2}(1+g_1^{-\frac{1}{2}})} - \int_0^{\frac{1}{2}(1-g_1^{-\frac{1}{2}})} + \int_{(p+1)/2p}^{g_1^{-\frac{1}{2}}}\right)\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},x\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}x \\
&= 2\sqrt{g_1}\int_{(p+1)/2p}^{1}\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},x\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}x + 1 - \sqrt{g_1} \\
&= 1-\sqrt{g_1}\left[\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{p+1}{2p}\right)}{p\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}+\frac{(p+1)^{\frac{1}{2}(p-1)}(p-1)^{\frac{p+1}{2}}}{2^p p^{p+1}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\right],
\end{aligned}
$$

using equations (3.36), (3.35), the expression for $\int_0^x \mathrm{B}(r,s,z)\mathrm{d}z$ and the fact that $\mathrm{B}(r,s,x) + \mathrm{B}(s,r,1-x) = \mathrm{B}(r,s)$. Hence $\Pr(C_r = 1)$ has the same form $\forall g_1 \in (0, 4p^2/(p+1)^2]$.

$g_1 \in [4p^2/(p+1)^2, p]$: Here the probability is

$$
\begin{aligned}
&\Pr(C_r=1) \\
&\quad = \int_0^{\sqrt{1-g_1/p}} \frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\nu-1}{\sqrt{g_1}}\right)\right)-\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\nu+1}{\sqrt{g_1}}\right)\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}\nu
\end{aligned}
$$

$$= 2\sqrt{g_1}\left(\int_{\frac{1}{2}(1+g_1^{-\frac{1}{2}}-(g_1^{-1}-p^{-1})^{\frac{1}{2}})}^{\frac{1}{2}(1+g_1^{-\frac{1}{2}})} - \int_{\frac{1}{2}(1-g_1^{-\frac{1}{2}}-(g_1^{-1}-p^{-1})^{\frac{1}{2}})}^{\frac{1}{2}(1-g_1^{-\frac{1}{2}})}\right)\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},x\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}x$$

$$= 2\sqrt{g_1}\int_{\frac{1}{2}(1+g_1^{-\frac{1}{2}}-(g_1^{-1}-p^{-1})^{\frac{1}{2}})}^{\frac{1}{2}(1+g_1^{-\frac{1}{2}}+(g_1^{-1}-p^{-1})^{\frac{1}{2}})}\frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},x\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\mathrm{d}x - \sqrt{\frac{p-g_1}{p}}$$

or

$$\begin{aligned}
&\Pr(C_r=1)\\
&= \frac{1}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)}\left[\left(1+\sqrt{\frac{p-g_1}{p}}\right)\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1+\frac{1}{\sqrt{g_1}}+\sqrt{\frac{p-g_1}{pg_1}}\right)\right)\right.\\
&\qquad -\left(1-\sqrt{\frac{p-g_1}{p}}\right)\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1+\frac{1}{\sqrt{g_1}}-\sqrt{\frac{p-g_1}{pg_1}}\right)\right)\\
&\qquad -\frac{\sqrt{g_1}}{2^p(p+1)}\left\{\left(\frac{p+1}{p}-\frac{2}{g_1}+\frac{2}{g_1}\sqrt{\frac{p-g_1}{p}}\right)^{\frac{p+1}{2}}\right.\\
&\qquad\qquad \left.\left.-\left(\frac{p+1}{p}-\frac{2}{g_1}-\frac{2}{g_1}\sqrt{\frac{p-g_1}{p}}\right)^{\frac{p+1}{2}}\right\}\right]\\
&\quad -\sqrt{\frac{p-g_1}{p}},
\end{aligned} \tag{3.38}$$

again using what was employed to derive the previous expression (all of these expressions could have been derived by integration in the $g_1\nu_1$-plane, but with no saving of effort).

$g_1 \in [p,\infty)$: Clearly, here $\Pr(C_r=1)=0$.

These results are illustrated in Figure 3.5.

We now wish to find the mean ratio of characteristic lengths, $\bar{C}_r$, and also the mean square ratio, $\overline{C_r^2}$. Clearly,

$$\bar{C}_r = \begin{cases} \left[\displaystyle\int_{\frac{4p^2}{(p+1)^2}}^{p}\left(\int_{(-\sqrt{g_1}+1)-}^{-\sqrt{1-\frac{g_1}{p}}} + \int_{\sqrt{1-\frac{g_1}{p}}}^{(\sqrt{g_1}-1)+}\right) + \int_{p}^{\infty}\int_{-\sqrt{g_1}+1}^{\sqrt{g_1}-1}\right] & \\ \qquad p_{\tilde{g}_1\tilde{\nu}_1}(g,\nu)C_r(g,\nu)\mathrm{d}\nu\mathrm{d}g + \Pr(C_r=1), & g_1\in(0,p]; \\ \displaystyle\int_{p}^{\infty}\int_{-\sqrt{g_1}+1}^{\sqrt{g_1}-1} p_{\tilde{g}_1\tilde{\nu}_1}(g,\nu)C_r(g,\nu)\mathrm{d}\nu\mathrm{d}g, & \text{otherwise,} \end{cases} \tag{3.39}$$

(as $\Pr(C_r=1)=0$ when $g\geq p$) that is, the probability that the characteristic length of the ellipsoid is unchanged plus the integral of C_r over the region where C_r is between zero and 1, weighted by the joint probability density of $\tilde{g}_1$ and $\tilde{\nu}_1$, given g_1.

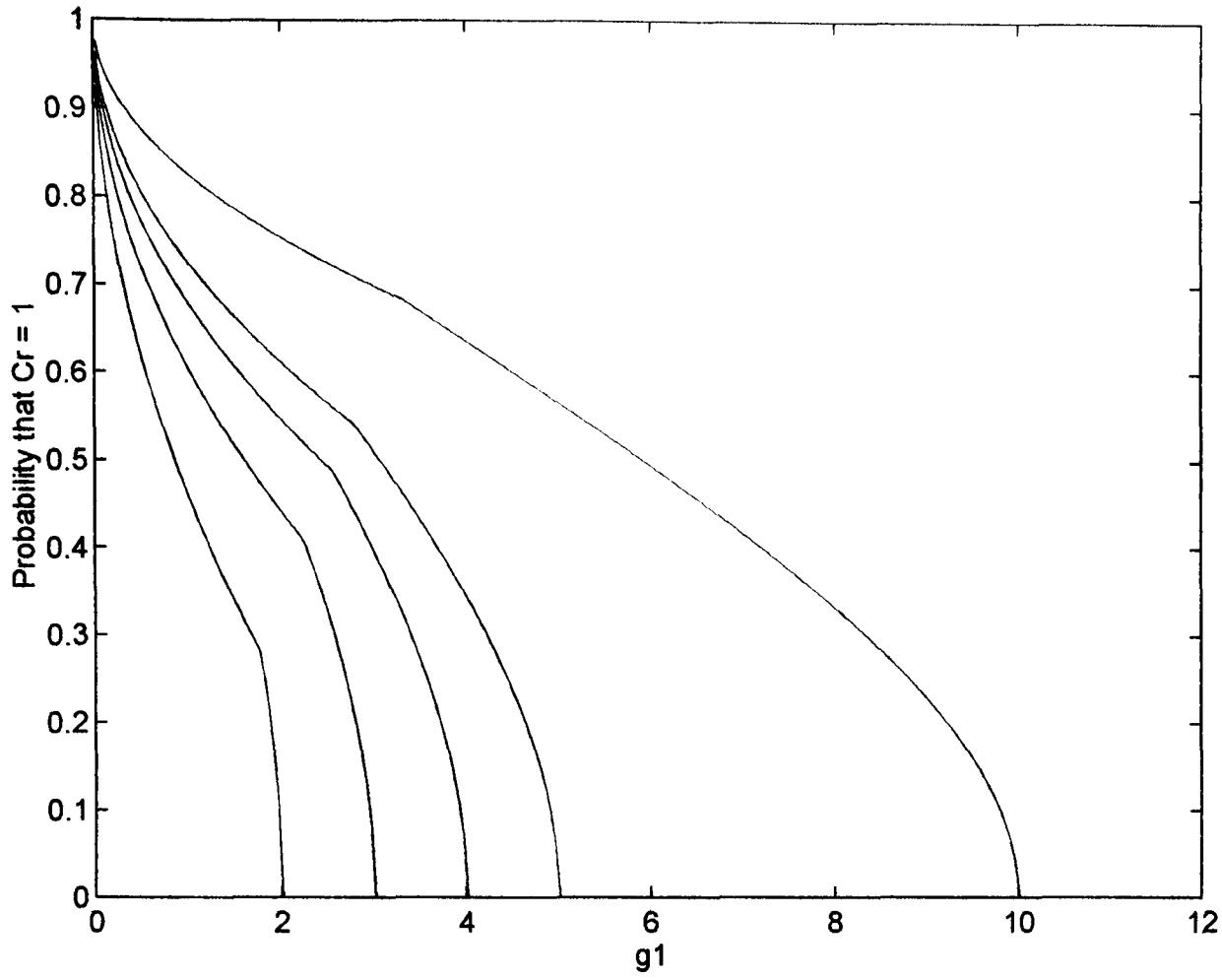


Figure 3.5: Probability that $C_r = 1$ (against g_1) (each curve reaches the horizontal axis at $g_1 = p$, the relevant dimension for the curve).

Looking at various ranges of g_1 separately, we have:

$g_1 \in (0, 1]$: Here, we can integrate equation (3.39) in terms of an infinite series of products of incomplete beta functions and elementary functions:

$$\begin{aligned}
&\bar{C}_r - \Pr(C_r = 1) \\
&= \sqrt{g_1} \int_{\frac{4p^2}{(p+1)^2}}^{\infty} g^{-\frac{3}{2}} \frac{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, g^{-\frac{1}{2}}\right)}{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} C_r\left(g, \sqrt{g} - 1\right) \mathrm{d}g \\
&= \sqrt{g_1} \frac{2p}{(p-1)^{\frac{p-1}{2p}} (p+1)^{\frac{p+1}{2p}}} \int_{\frac{4p^2}{(p+1)^2}}^{\infty} \frac{\left(\sqrt{g} - 1\right)^{\frac{p-1}{2p}}}{\sqrt{g}} g^{-\frac{3}{2}} \frac{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, g^{-\frac{1}{2}}\right)}{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} \mathrm{d}g \\
&= \sqrt{g_1} \frac{4p}{(p-1)^{\frac{p-1}{2p}} (p+1)^{\frac{p+1}{2p}} \mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} \\
&\qquad \times \int_0^{\frac{(p+1)}{2p}} x^{\frac{p+1}{2p}} (1-x)^{\frac{p-1}{2p}} \mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, x\right) \mathrm{d}x. \qquad (3.40)
\end{aligned}$$

This can now be integrated in terms of an infinite series:

$$\int_0^{\frac{p+1}{2p}} x^{\frac{p+1}{2p}}(1-x)^{\frac{p-1}{2p}}\mathrm{B}\left(\tfrac{p+1}{2},\tfrac{p+1}{2},x\right)\mathrm{d}x$$

$$= \sum_{k=0}^{\infty}\frac{\left(-\frac{p-1}{2p},k\right)}{k!}\int_0^{\frac{p+1}{2p}} x^{\frac{p+1}{2p}+k}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},x\right)\mathrm{d}x$$

(by the fact that the series given by the binomial theorem for $(1-x)^{\frac{p-1}{2p}}$ is uniformly convergent on $[0,(p+1)/2p]$, allowing the interchange of integration and summation)

$$= \sum_{k=0}^{\infty}\frac{\left(-\frac{p-1}{2p},k\right)}{k!\left(\frac{3p+1}{2p}+k\right)}\left[\left(\frac{p+1}{2p}\right)^{\frac{3p+1}{2p}+k}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{p+1}{2p}\right)\right.$$

$$\left.-\,\mathrm{B}\left(\frac{p^2+4p+1}{2p}+k,\frac{p+1}{2},\frac{p+1}{2p}\right)\right]$$

where $(\cdot,\cdot)$ is Appell's symbol

$$(a,n) := \begin{cases} 1, & n=0; \\ a(a+1)\cdots(a+n-1) = (a+n-1)(a,n-1), & n\in\mathbb{N}_+ := \{1,2,\ldots\}; \\ \frac{1}{(a-1)(a-2)\cdots(a+n)} = (a,n+1)/(a+n), & -n\in\mathbb{N}_+ \end{cases} \tag{3.41}$$

(see, for example, [5]).

Since $\mathrm{B}(r,s,x)=\int_0^x t^{r-1}(1-t)^{s-1}\mathrm{d}t \le \int_0^x t^{r-1}\mathrm{d}t = \frac{x^r}{r}$,

$$\left|\sum_{k=r+1}^{\infty}\frac{\left(-\frac{p-1}{2p},k\right)}{k!\left(\frac{3p+1}{2p}+k\right)}\left[\left(\frac{p+1}{2p}\right)^{\frac{3p+1}{2p}+k}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{p+1}{2p}\right)\right.\right.$$

$$\left.\left.-\,\mathrm{B}\left(\frac{p^2+4p+1}{2p}+k,\frac{p+1}{2},\frac{p+1}{2p}\right)\right]\right| \le$$

$$\sum_{k=r+1}^{\infty}\left|\frac{\left(-\frac{p+1}{2p},k\right)}{k!\left(\frac{3p+1}{2p}+k\right)}\right|\left(\frac{p+1}{2p}\right)^{\frac{3p+1}{2p}+k}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{p+1}{2p}\right) \le$$

$$\sum_{k=r+1}^{\infty}\left(\frac{p+1}{2p}\right)^{\frac{3p+1}{2p}+k}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{p+1}{2p}\right) \le$$

$$\sum_{k=r+1}^{\infty}\frac{2}{p+1}\left(\frac{p+1}{2p}\right)^{\frac{p^2+4p+1}{2p}+k} =$$

$$\frac{2}{p-1}\left(\frac{p+1}{2p}\right)^{\frac{p^2+2(r+2)p+1}{2p}} \tag{3.42}$$

for positive integral r, so the effect of truncating series (3.41) after r terms is no worse than truncating a geometric series with initial term $\frac{2}{p+1}\left(\frac{p+1}{2p}\right)^{\frac{p^2+4p+1}{2}}$ and common ratio $\frac{p+1}{2p}$ after the same number of terms.

Now we can write equation (3.40) as

$$\bar{C}_r = 1 - K_C\sqrt{g_1}, \tag{3.43}$$

where

$$\begin{aligned}
&\mathrm{B}\left(\tfrac{p+1}{2},\tfrac{p+1}{2}\right)K_C \\
&= \left[\frac{1}{p} - \frac{4p}{(p-1)^{\frac{p-1}{2p}}(p+1)^{\frac{p+1}{2p}}}\sum_{k=0}^{\infty}\frac{\left(-\frac{p-1}{2p},k\right)}{\left(\frac{3p+1}{2p}+k\right)k!}\left(\frac{p+1}{2p}\right)^{\frac{3p+1}{2p}+k}\right]\mathrm{B}\left(\tfrac{p+1}{2},\tfrac{p+1}{2},\tfrac{p+1}{2p}\right) \\
&\quad + \frac{4p}{(p-1)^{\frac{p-1}{2p}}(p+1)^{\frac{p+1}{2p}}}\sum_{k=0}^{\infty}\frac{\left(-\frac{p-1}{2p},k\right)}{k!\left(\frac{3p+1}{2p}+k\right)}\mathrm{B}\left(\tfrac{p^2+4p+1}{2p}+k,\tfrac{p+1}{2},\tfrac{p+1}{2p}\right) \\
&\quad + \frac{(p-1)^{\frac{p+1}{2}}(p+1)^{\frac{1}{2}(p-1)}}{2^p p^{p+1}}
\end{aligned} \tag{3.44}$$

A comparison of $1-K_C$, which is the mean value of C_r when $g_1 = 1$, with $\Pr(C_r = 1)$ when $g_1 = 1$, for various values of p, is made in Figure (3.6), where the bounds of expression (3.42) have been used to ensure that K_C is accurate to four significant places.

$g_1 \in [1, 4p^2/(p+1)^2]$: Here

$$\bar{C}_r - \Pr(C_r = 1) = 2\int_{4p^2/(p+1)^2}^{\infty} \bar{p}_{\tilde{g}_1\tilde{\nu}_1}(g, \sqrt{g}-1)C_r(g,\sqrt{g}-1)\mathrm{d}g$$

where $\bar{p}_{\tilde{g}_1\tilde{\nu}_1}$ has the same form as $p_{\tilde{g}_1\tilde{\nu}_1}$ but with any δ "functions" deleted (as they have been integrated out), so $\bar{C}_r - \Pr(C_r = 1)$ has the same form as for $g_1 \in (0,1]$, and, as $\Pr(C_r = 1)$ is also unaltered, we again have expression (3.43) with K_C given by equation (3.44) for $\bar{C}_r$.

$g_1 \in [4p^2/(p+1)^2, p]$: Now

$$\begin{aligned}
&\bar{C}_r - \Pr(C_r = 1) \\
&\quad = 2\int_{\sqrt{1-g_1/p}}^{\sqrt{g_1}-1} \bar{p}_{\tilde{g}_1\tilde{\nu}_1}(g_1,\nu)C_r(g_1,\nu)\mathrm{d}\nu + 2\int_{g_1}^{\infty}\bar{p}_{\tilde{g}_1\tilde{\nu}_1}(g,\sqrt{g}-1)C_r(g,\sqrt{g}-1)\mathrm{d}g.
\end{aligned}$$

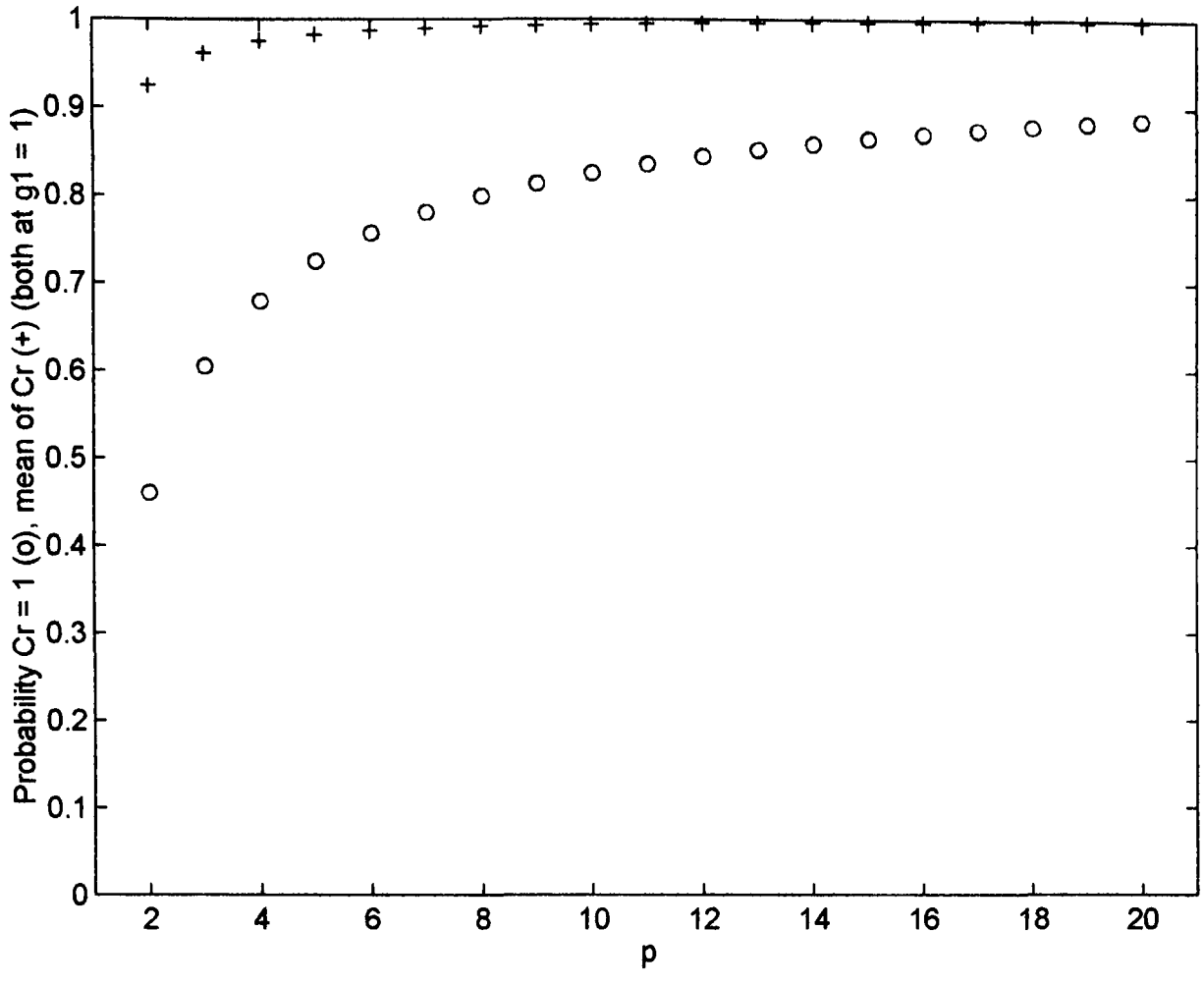


Figure 3.6: Comparison of $\bar{C}_r|_{g_1=1} = 1 - K_C$ with $\Pr(C_r = 1)|_{g_1=1}$

The first term of this is

$$\int_{\sqrt{1-g_1/p}}^{\sqrt{g_1}-1} \frac{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, \frac{1}{2}\left(1 - \frac{\nu-1}{\sqrt{g_1}}\right)\right) - \mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\left(1 - \frac{\nu+1}{\sqrt{g_1}}\right)\right)}{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} C_r(g_1, \nu)\mathrm{d}\nu := C_1(g_1), \text{ say,} \tag{3.45}$$

and the second is

$$\sqrt{g_1}\frac{4p}{(p-1)^{\frac{p-1}{2p}}(p+1)^{\frac{p+1}{2p}}}\int_0^{1/\sqrt{g_1}} x^{\frac{p+1}{2p}}(1-x)^{\frac{p-1}{2p}}\frac{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, x\right)}{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)}\mathrm{d}x =$$

$$\frac{4p\sqrt{g_1}}{(p-1)^{\frac{p-1}{2p}}(p+1)^{\frac{p+1}{2p}}\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)}$$

$$\times \sum_{k=0}^{\infty} \frac{\left(-\frac{p-1}{2p}, k\right)}{k!\left(\frac{3p+1}{2p} + k\right)}\left[g_1^{-\frac{(3+2k)p+1}{4p}}\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, g_1^{-\frac{1}{2}}\right) - \mathrm{B}\left(\frac{p^2+4p+1}{2p} + k, \frac{p+1}{2}, g_1^{-\frac{1}{2}}\right)\right] :=$$

$$C_2(g_1), \text{ say.} \tag{3.46}$$

The bound on the truncation error, analogous to equation (3.42), is

$$\left| \sum_{k=r+1}^{\infty} \frac{\left(-\frac{p-1}{2p}, k\right)}{k!\left(\frac{3p+1}{2p}+k\right)} \left[g_1^{-\frac{(3+2k)p+1}{4p}} \mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, g_1^{-\frac{1}{2}}\right) \right.\right.$$

$$\left.\left. - \mathrm{B}\left(\frac{p^2+4p+1}{2p}+k, \frac{p+1}{2}, g_1^{-\frac{1}{2}}\right) \right] \right| \leq$$

$$\sum_{k=r+1}^{\infty} \frac{2}{p+1} g_1^{-\frac{p^2+4p+1+2k}{4p}} =$$

$$\frac{2g_1^{-\frac{p^2+2(r+3)p+1}{4p}}}{(p+1)(1-g_1^{-\frac{1}{2}})} \tag{3.47}$$

We evaluate $C_1(g_1)$ by numerical integration (Simpson's rule suffices) and then use

$$\bar{C}_r(g_1) = \Pr(C_r = 1) + C_1(g_1) + C_2(g_1) \tag{3.48}$$

to find $\bar{C}_r(g_1)$.

$g_1 \in [p, \infty)$: Here $\Pr(C_r = 1) = 0$, and

$$\bar{C}_r(g_1) = C_1(g_1) + C_2(g_1) \tag{3.49}$$

where

$$C_1(g_1) = \int_0^{\sqrt{g_1}-1} \frac{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, \frac{1}{2}\left(1-\frac{\nu-1}{\sqrt{g_1}}\right)\right) - \mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\left(1-\frac{\nu+1}{\sqrt{g_1}}\right)\right)}{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} C_r(g_1, \nu)\mathrm{d}\nu, \tag{3.50}$$

and $C_2(g_1)$ is again given by equation (3.46).

These results for $\bar{C}_r(g_1)$ are illustrated in Figure (3.7).

By similar arguments to those involved in finding $\bar{C}_r$, it is found that when:

$g_1 \in (0, 4p^2/(p+1)^2]$:

$$\overline{C_r^2}(g_1) = 1 - K_{C^2}\sqrt{g_1}, \tag{3.51}$$

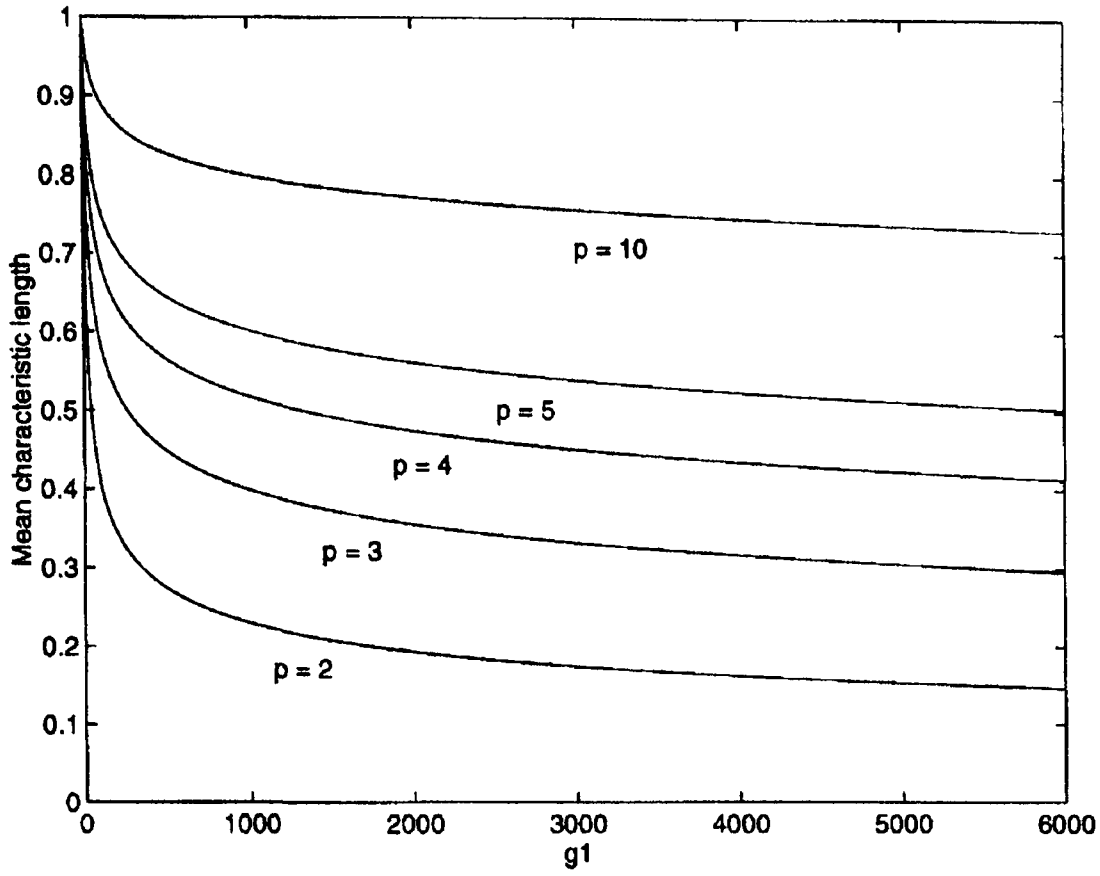


Figure 3.7: Mean Value of C_r against g_1.

where

$$\mathrm{B}\left(\tfrac{p+1}{2},\tfrac{p+1}{2}\right) K_{C^2}$$

$$= \left[\frac{1}{p} - \frac{8p^2}{(p-1)^{\frac{p-1}{p}}(p+1)^{\frac{p+1}{p}}} \sum_{k=0}^{\infty} \frac{\left(-\frac{p-1}{p},k\right)}{\left(\frac{2p+1}{p}+k\right)k!}\left(\frac{p+1}{2p}\right)^{\frac{2p+1}{p}+k}\right] \mathrm{B}\left(\tfrac{p+1}{2},\tfrac{p+1}{2},\tfrac{p+1}{2p}\right)$$

$$+ \frac{8p^2}{(p-1)^{\frac{p-1}{p}}(p+1)^{\frac{p+1}{p}}} \sum_{k=0}^{\infty} \frac{\left(-\frac{p-1}{p},k\right)}{k!\left(\frac{2p+1}{p}+k\right)} \mathrm{B}\left(\tfrac{p^2+5p+2}{2p}+k,\tfrac{p+1}{2},\tfrac{p+1}{2p}\right)$$

$$+ \frac{(p-1)^{\frac{p+1}{2}}(p+1)^{\frac{1}{2}(p-1)}}{2^p p^{p+1}} \tag{3.52}$$

and the effect of truncation is given by

$$\left| \sum_{k=r+1}^{\infty} \frac{\left(-\frac{p-1}{p},k\right)}{k!\left(\frac{2p+1}{p}+k\right)} \left[\left(\frac{p+1}{2p}\right)^{\frac{2p+1}{p}+k} \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{p+1}{2p}\right) - \mathrm{B}\left(\frac{p^2+5p+2}{2p}+k,\frac{p+1}{2},\frac{p+1}{2p}\right)\right]\right| =$$

$$\frac{2}{p-1}\left(\frac{p+1}{2p}\right)^{\frac{p^2+(2r+5)p+2}{2p}} \tag{3.53}$$

A graphical comparison of $1 - K_{C^2}$, the square mean value of C_r when $g_1 = 1$, with

$\Pr(C_r = 1)$ when $g_1 = 1$, is made in Figure (3.8), where the bounds of expression (3.53) have been used to ensure accuracy.

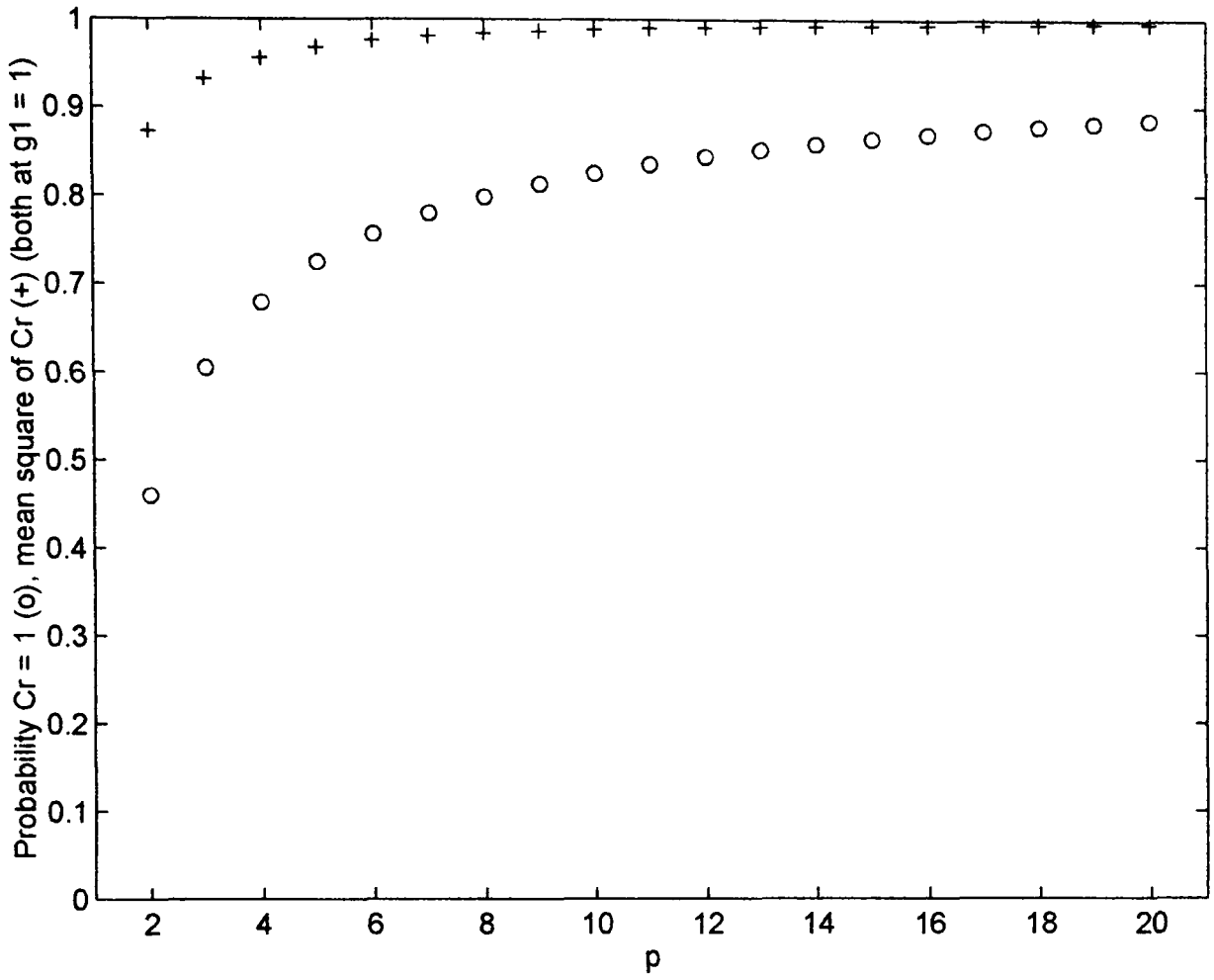


Figure 3.8: Comparison of $\overline{C_r^2}|_{g_1=1} = 1 - K_{C^2}$ with $\Pr(C_r = 1)|_{g_1=1}$

$g_1 \in [4p^2/(p+1)^2, p]$: Now

$$\overline{C_r^2} - \Pr(C_r = 1) = C_{21}(g_1) + C_{22}(g_1), \text{ say,} \tag{3.54}$$

where

$$C_{12}(g_1) = \int_{\sqrt{1-g_1/p}}^{\sqrt{g_1}-1} \frac{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}, \frac{1}{2}\left(1 - \frac{\nu-1}{\sqrt{g_1}}\right)\right) - \mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\left(1 - \frac{\nu+1}{\sqrt{g_1}}\right)\right)}{\mathrm{B}\left(\frac{p+1}{2}, \frac{p+1}{2}\right)} C_r(g_1, \nu)^2 \mathrm{d}\nu \tag{3.55}$$

and

$$C_{22}(g_1) = \frac{8p^2\sqrt{g_1}}{(p-1)^{\frac{p-1}{p}}(p+1)^{\frac{p+1}{p}}\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)} \times \sum_{k=0}^{\infty} \frac{\left(-\frac{p-1}{p},k\right)}{k!\left(\frac{2p+1}{p}+k\right)} \left[g_1^{-\frac{(k+2)p+1}{2p}} \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},g_1^{-\frac{1}{2}}\right) - \mathrm{B}\left(\frac{p^2+5p+2}{2p}+k,\frac{p+1}{2},g_1^{-\frac{1}{2}}\right)\right]. \tag{3.56}$$

A bound on the truncation error is given by

$$\left|\sum_{k=0}^{\infty} \frac{\left(-\frac{p-1}{p},k\right)}{k!\left(\frac{2p+1}{p}+k\right)} \left[g_1^{-\frac{(k+2)p+1}{2p}} \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},g_1^{-\frac{1}{2}}\right) - \mathrm{B}\left(\frac{p^2+5p+2}{2p}+k,\frac{p+1}{2},g_1^{-\frac{1}{2}}\right)\right]\right| \leq \frac{2g_1^{-\frac{p^2+2(r+7)p+2}{4p}}}{(p+1)(1-g_1^{-\frac{1}{2}})}. \tag{3.57}$$

$g_1 \in [p,\infty)$: Here

$$\overline{C_r^2} = C_{21}(g_1) + C_{22}(g_1), \text{ say,} \tag{3.58}$$

where C_{21} is given by

$$C_{12}(g_1) = \int_0^{\sqrt{g_1}-1} \frac{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2},\frac{1}{2}\left(1-\frac{\nu-1}{\sqrt{g_1}}\right)\right) - \mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\left(1-\frac{\nu+1}{\sqrt{g_1}}\right)\right)}{\mathrm{B}\left(\frac{p+1}{2},\frac{p+1}{2}\right)} C_r(g_1,\nu)^2 \mathrm{d}\nu, \tag{3.59}$$

and C_{22} is still given by equation (3.56).

These results are displayed in Figure 3.9.

3.3 Empirical Results for the Fogel-Huang Algorithm

3.3.1 Batch-optimal Results

In Figures 3.11, 3.12 and 3.13, we show the characteristic lengths of the minimum-volume ellipsoids containing the vertices of the polytopes formed from the intersection of the hyperplanes encountered in the processing of the Fogel-Huang algorithm up to the given step. The aim of displaying these volumes is to provide a bench-mark for the performance of the Fogel-Huang and other algorithms.

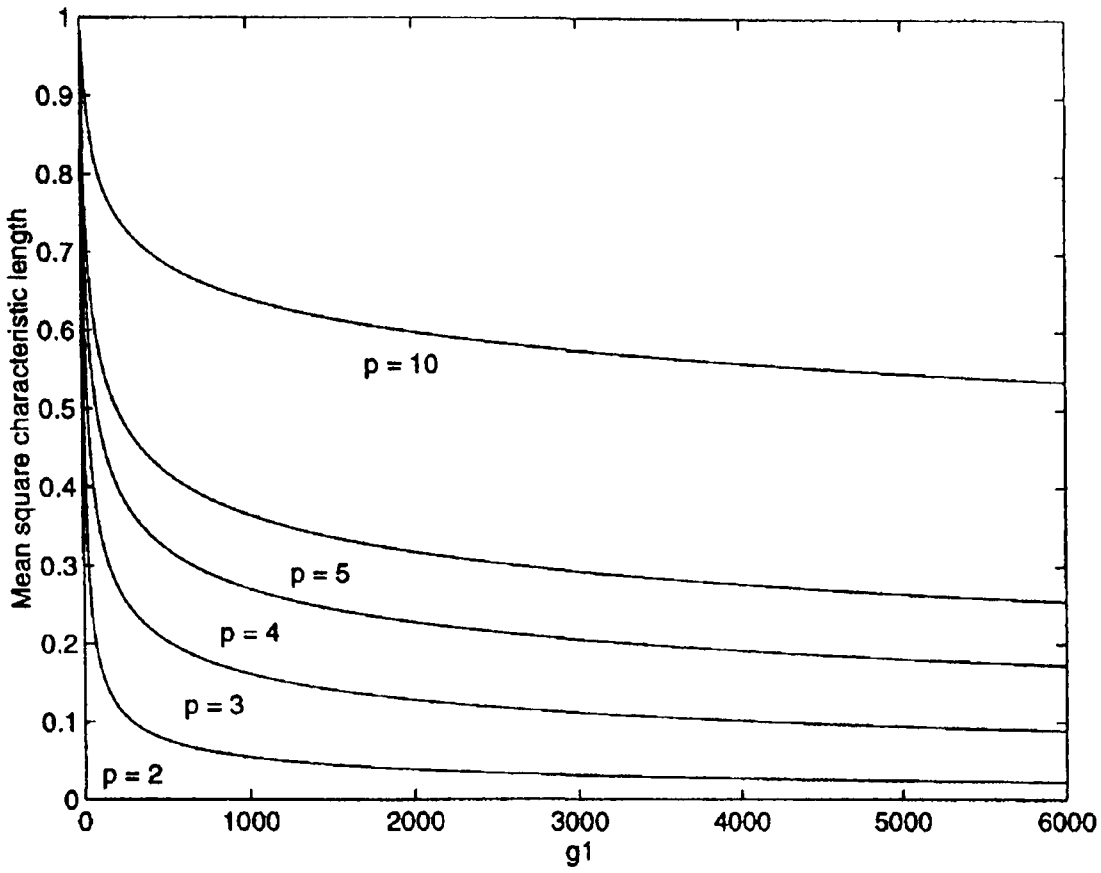


Figure 3.9: Mean Value of C_r^2 against g_1.

Figure 3.9: Mean Value of C_r^2 against g_1.

Obviously, when the number of hyperplane pairs encountered is less than the dimension of the space, the polytope defining the feasible set will not be closed (and it is not necessarily closed even if the number of hyperplanes is greater than the dimension, although the probability of it being unbounded is zero in principle, and almost zero even with finite computer precision), so we augment the hyperplane pairs with the p pairs which form the p-cube containing our initial ellipsoid (as shown in Figure 3.10). This will mean that the initial characteristic length of these minimum-volume ellipsoids (the pth root of the volume of the minimum-volume ellipsoid containing the p-cube containing our initial ellipsoid) will be greater (by a factor of $\sqrt{p}$) than the initial characteristic length of our Fogel-Huang ellipsoids (the pth root of the volume of our initial ellipsoid), and it is possible (although unlikely) that the reference characteristic length will exceed the Fogel-Huang characteristic length up to the $(p-1)$st step.

These minimum-volume ellipsoids are obtained by first finding the vertices of the polytope defining the posterior feasible set at the particular step. This is done by satisfying the equalities derived from the inequalities satisfied by the feasible set p at a time and then checking that the remaining such equalities are satisfied. Then the minimum-volume ellipsoid containing these vertices is found using the algorithm presented by Pronzato and Walter[22].

In the Figures 3.11 to 3.13 (and in subsequent Figures showing the same kind of information),

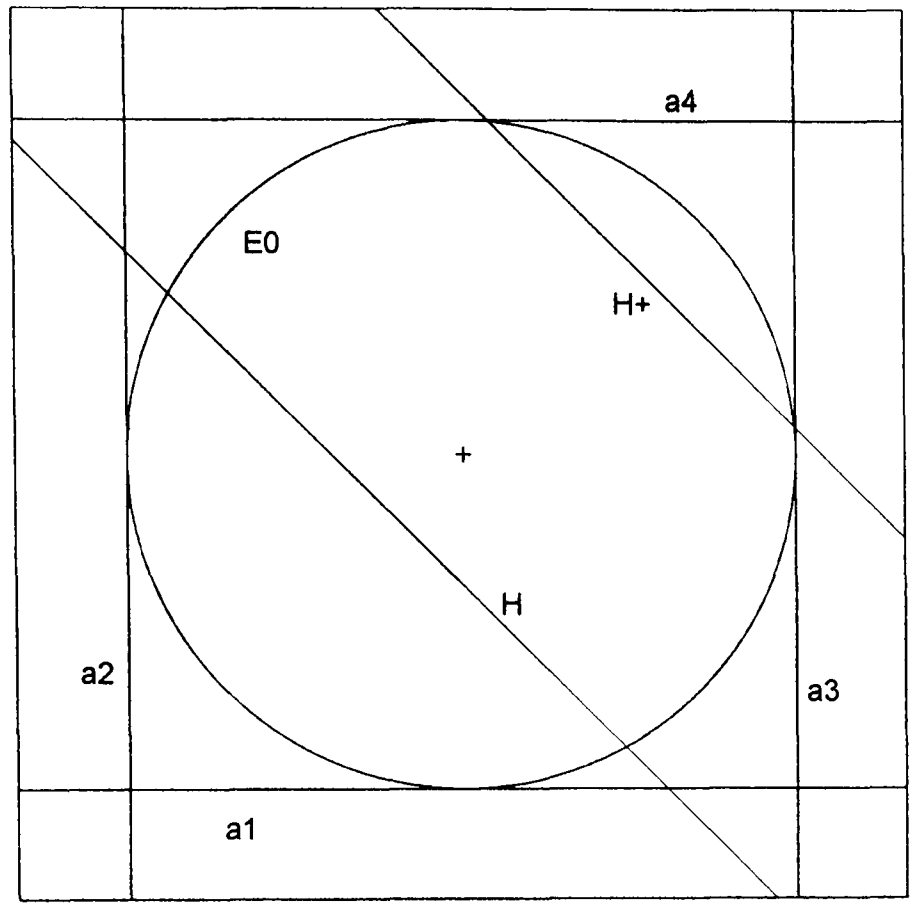


Figure 3.10: Augmenting hyperplanes for the calculation of the earlier minimum-volume ellipsoids (a*i*: augmenting hyperplanes, E0: initial ellipsoid, H±: first hyperplane pair)

each dot represents the characteristic length associated with a particular data set, and the solid lines are the maximum, mean and minimum of these lengths over all 100 data sets.

As might be expected, the characteristic length, in the majority of cases, diminishes less rapidly after the pth hyperplane pair is brought in, as the polytope formed from the first p hyperplanes will almost always be bounded and closer in volume to those formed from more than p hyperplane pairs, and the ratio of the volumes of two convex polytopes will tend to be reflected in the ratio of the volumes of the minimum-volume ellipsoids containing them.

If plots corresponding to the same dimension in parameter space but with data affected by noise with different distributions are compared, we can see that the mean characteristic length increases as the standard deviation of the noise affecting the data decreases. This might also have been expected, as data further away from the bounds is less informative, and the smaller the dispersion caused by noise with a lower standard deviation, the further away from the bounds the data will be on average.

This effect is partially obscured by the greater spread of the data with the higher standard deviation, as this often results in a greater difference between the maximum and minimum characteristic lengths at a given step number.

It is also observed that the distribution of the characteristic lengths about the mean is skewed,

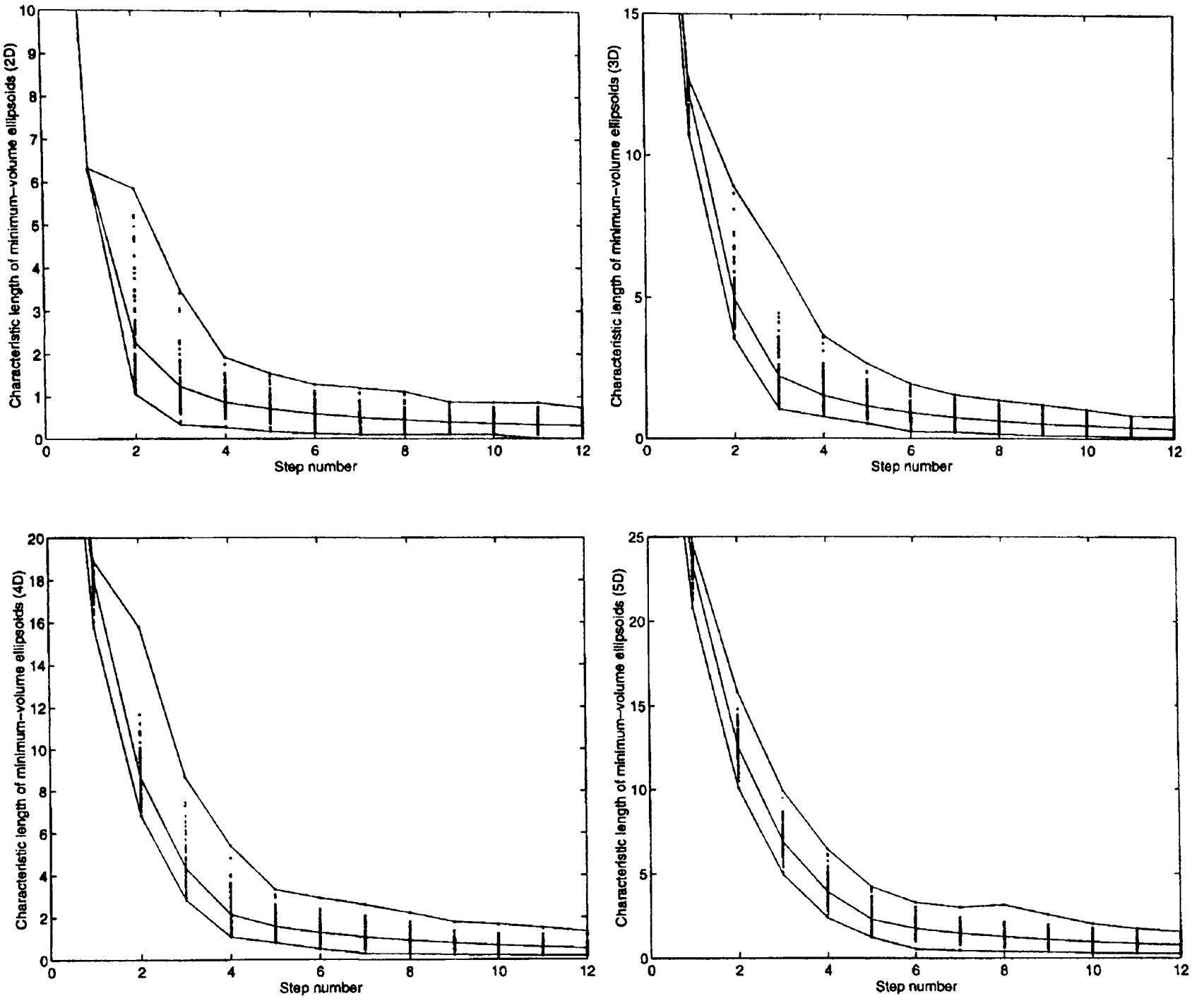
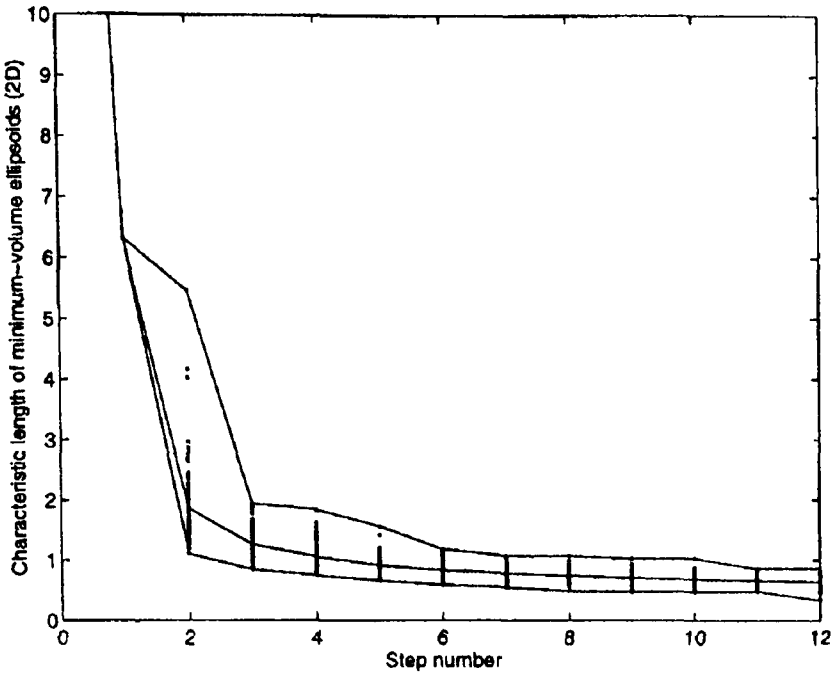


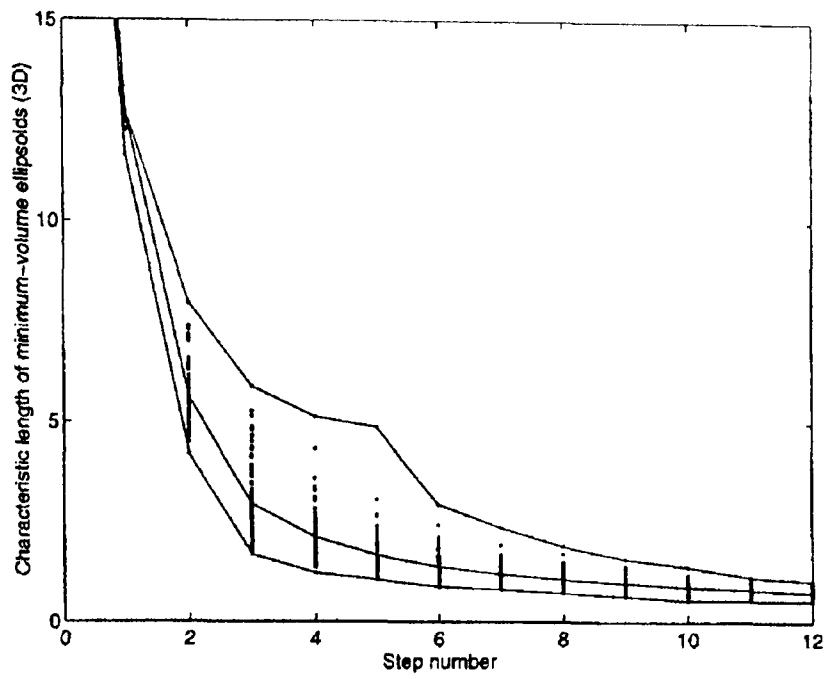
Figure 3.11: Characteristic lengths of BLJ ellipsoids (noise uniformly distributed). Clockwise from the upper right-hand side, the figures are for 2-, 3-, 4- and 5-dimensional parameter space, and the initial values of the characteristic lengths are $20\sqrt{2}$, $20\sqrt{3}$, 40 and $20\sqrt{5}$.

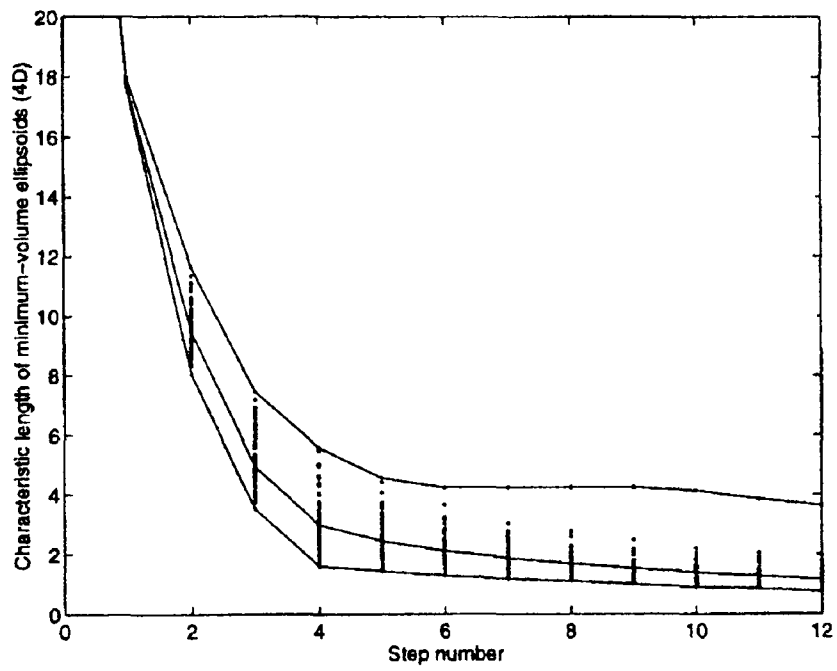
so that the performance is particularly bad for a few exceptional sets. This effect is more prominent for early step numbers (corresponding to less than p hyperplane pairs which do not, by themselves, define bounded polytopes), but persists up to the final step.

Figures 3.14 to 3.13 are similar to the Figures 3.11 to 3.13, but they display the distance from the centre of the minimum-volume ellipsoid to the true parameter value. Thus these Figures help assess the value of the centre of the minimum-volume ellipsoid as a point estimate of the parameter value.

As this centre-parameter distance is not directly optimised, it would be expected to increase as well as decrease during a run, although the decrements would be expected, on average, to outweigh the increments as the polytope boundaries come closer to the parameter value. This







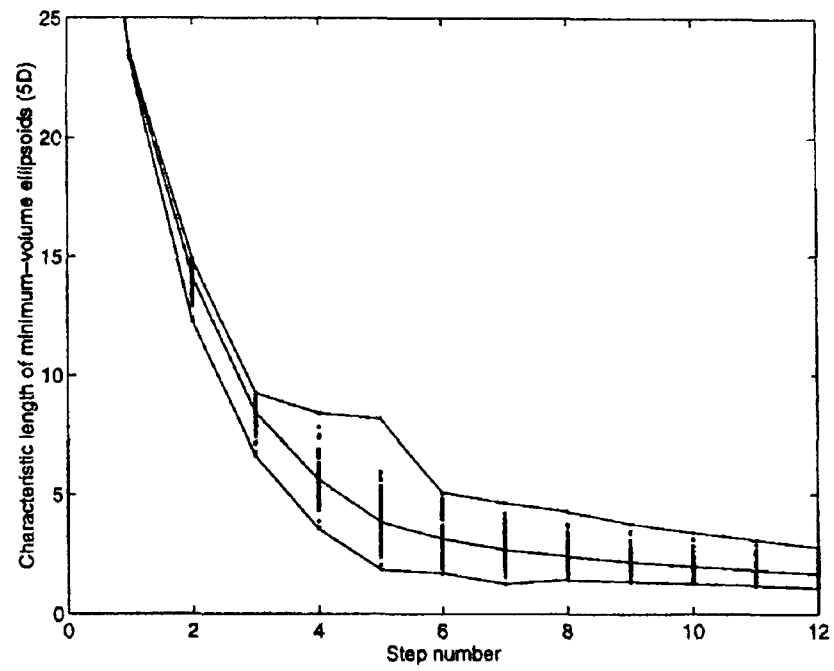


Figure 3.12: Characteristic lengths of minimum-volume ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

is indeed the observed mean behaviour.

The massive skew in these results means that a few extremely bad cases have a great effect on the average performance. Geometrically, these bad cases are probably those where the true parameter lies close to a vertex of the enclosing polytope and, in addition the vertex is situated on the minimum-volume ellipsoid close to a maximal semi axis.

Also, it is observed that the greater the standard deviation of the noise affecting the data, the greater the mean centre-parameter distance. This seems to be due to the fact the greater the spread of the data, the more likely it is that one hyperplane from a pair will lie close to the true parameter value, and that the greater the difference between the distances from the members of a pair of hyperplanes to the true parameter value is, the more the centre of the resulting

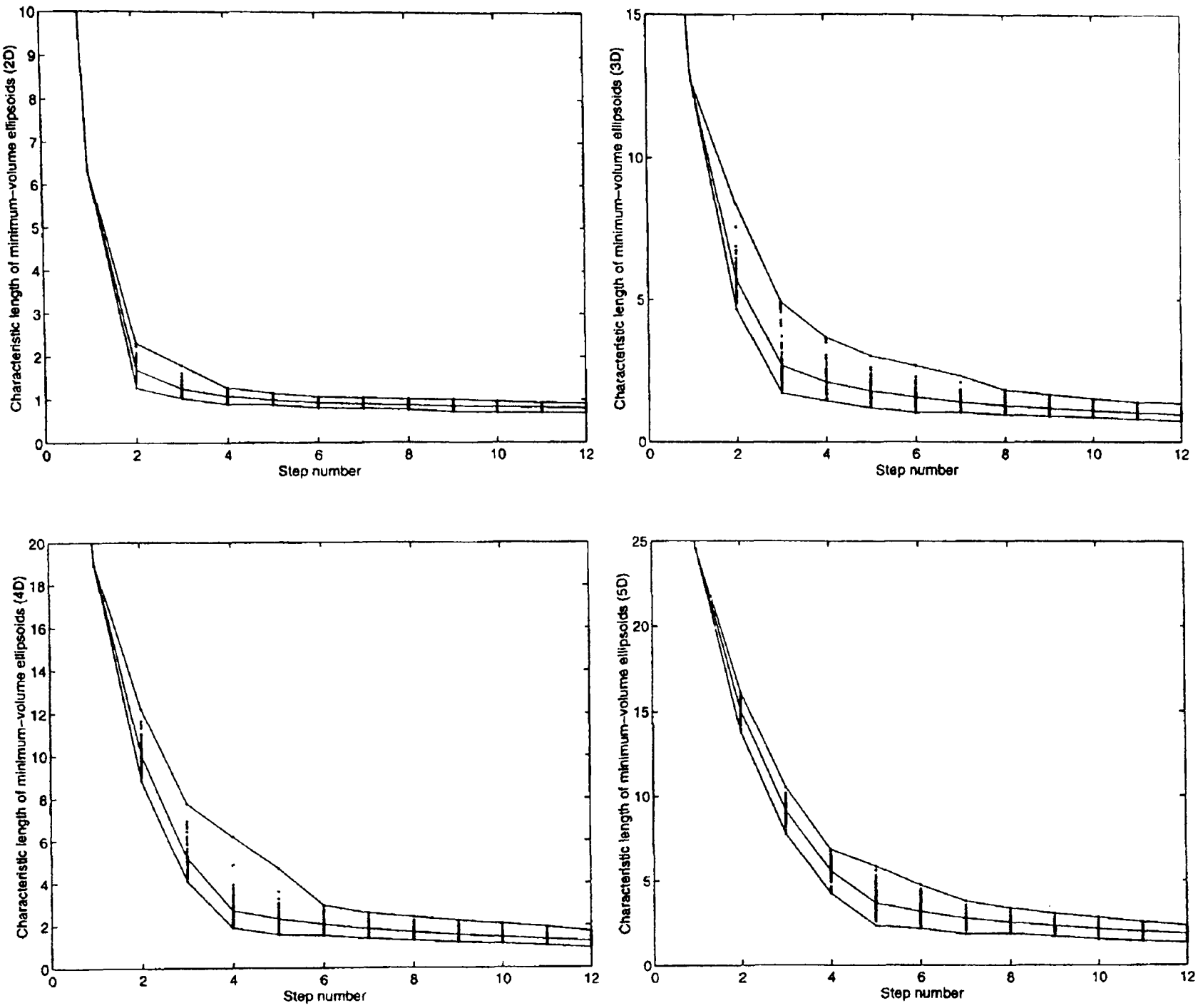


Figure 3.13: Characteristic lengths of minimum-volume ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

ellipsoid is biased away from the true value.

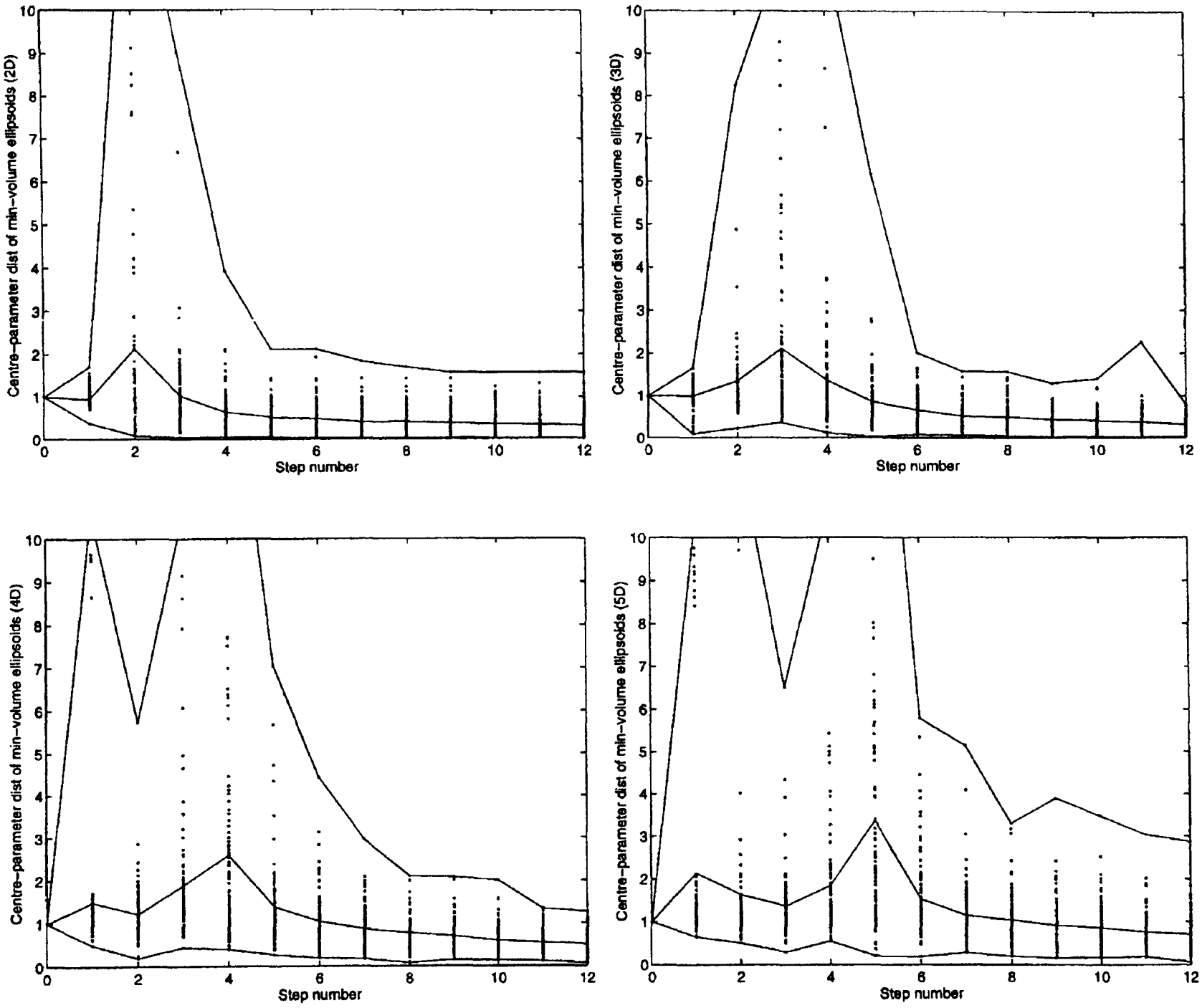


Figure 3.14: Ellipsoid centre to true parameter distance of minimum- volume ellipsoids (noise uniformly distributed).

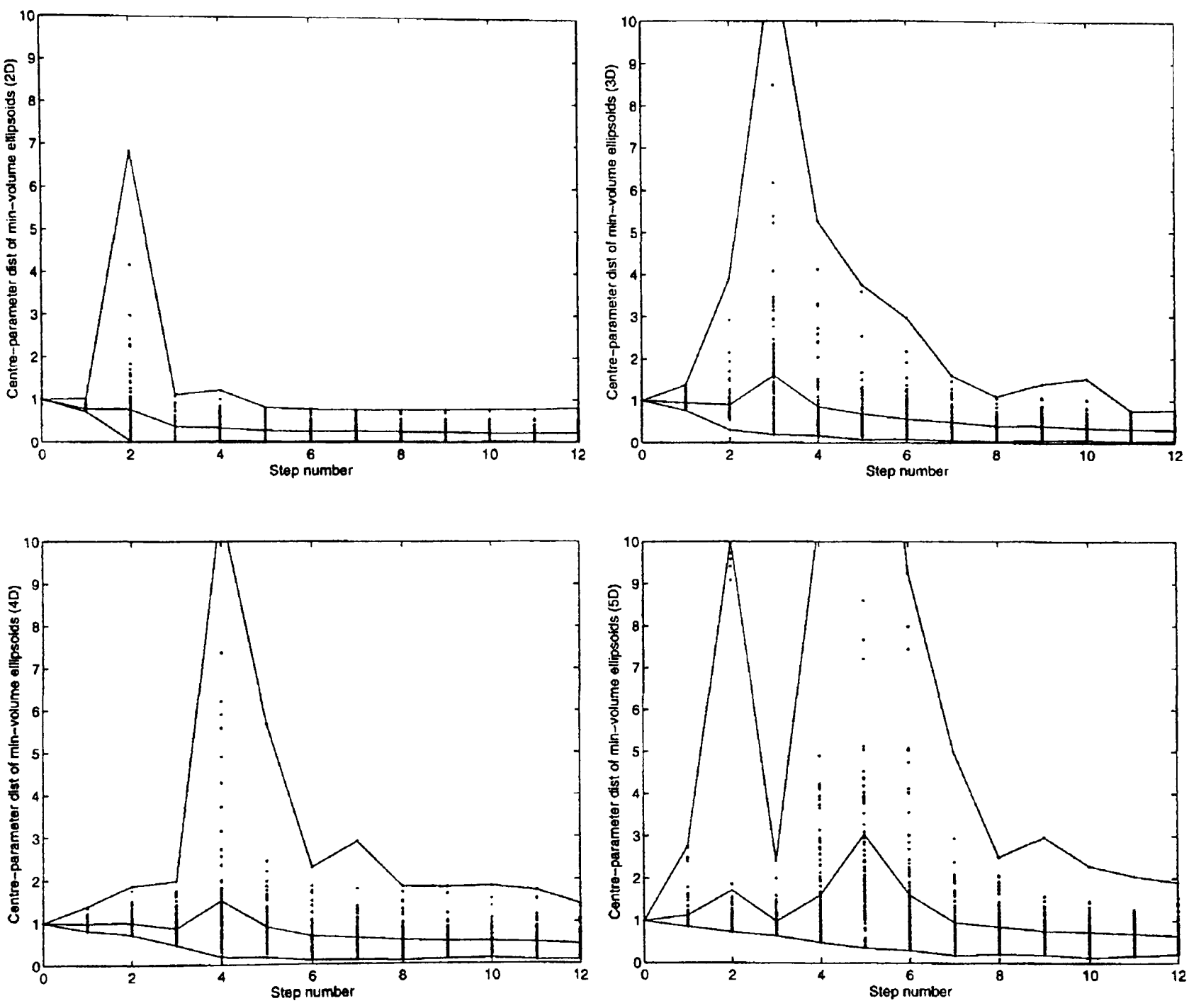


Figure 3.15: Ellipsoid centre to true parameter distance of minimum-volume ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

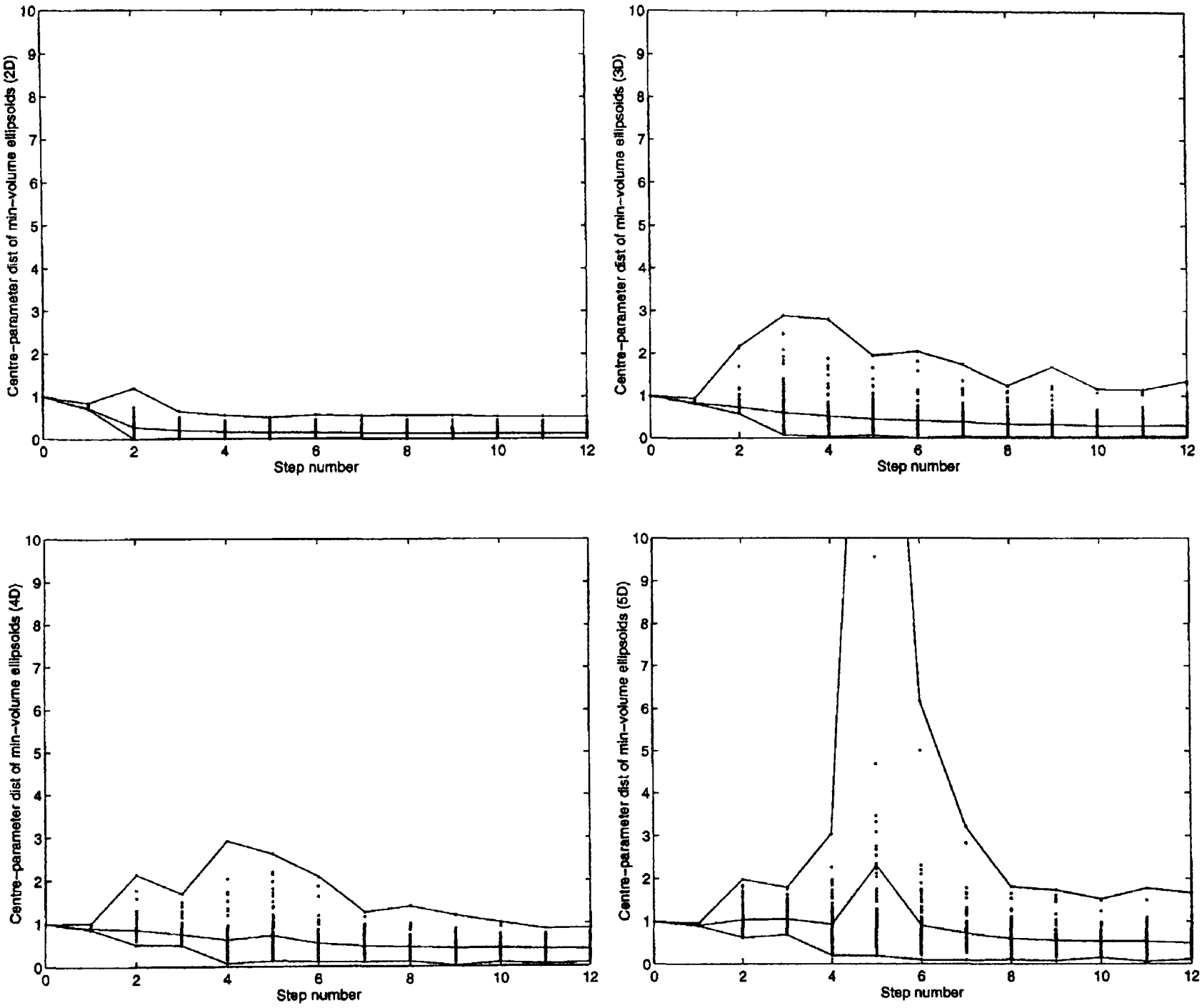


Figure 3.16: Ellipsoid centre to true parameter distance of minimum-volume ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

3.3.2 The Fogel-Huang Algorithm Without Data Recycling

Figures 3.17 to 3.19 are similar to Figures 3.11 to 3.13, but now the ellipsoids are derived from the Fogel-Huang algorithm, and for each data set the characteristic length of the Fogel-Huang ellipsoid is divided by the characteristic length of the minimum-volume ellipsoid about the final feasible parameter set; i.e., the polytope bounded by the nonredundant members of all 12 pairs of hyperplanes, as found in subsection 3.3.1.

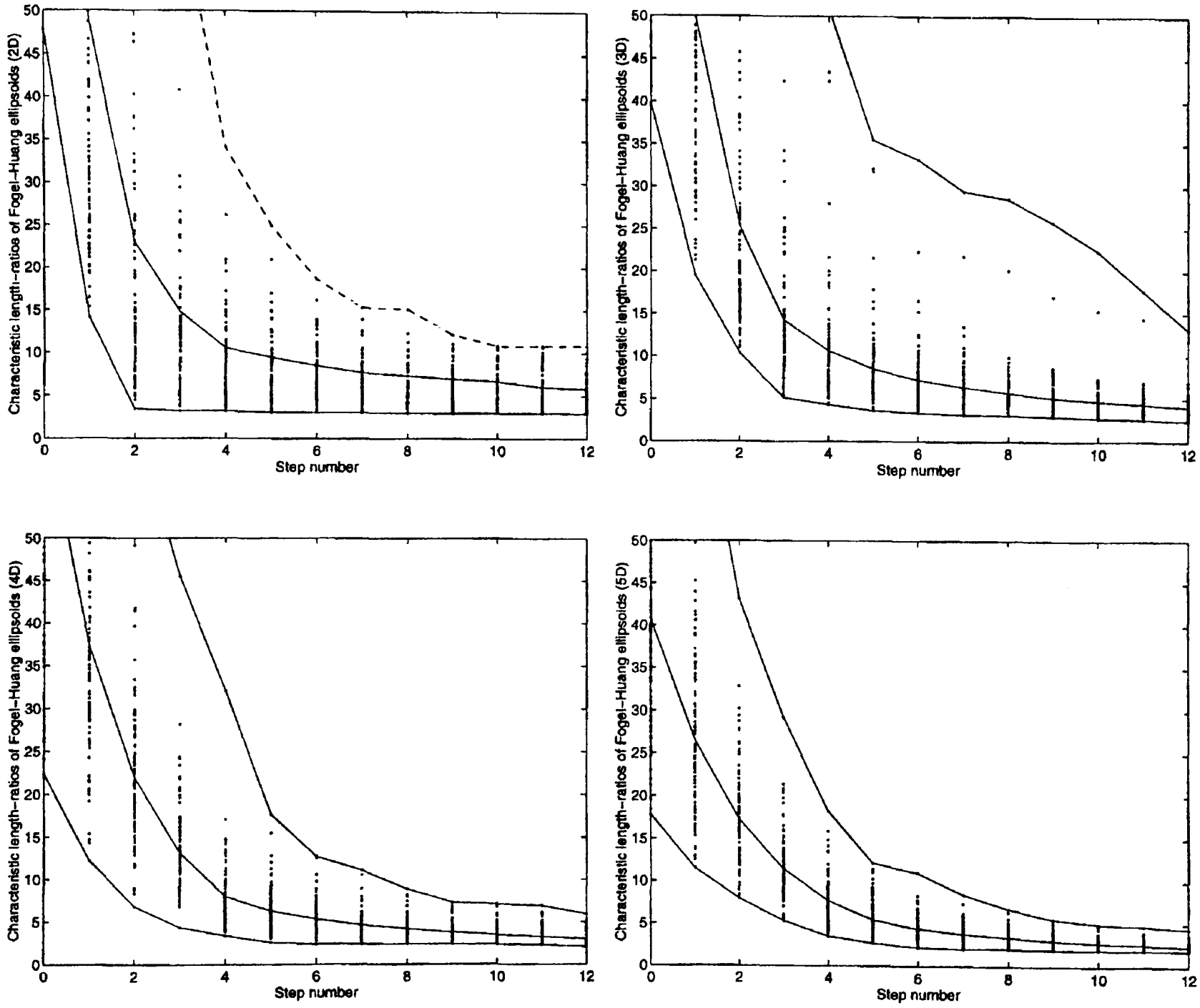


Figure 3.17: Characteristic lengths of Fogel-Huang ellipsoids (noise uniformly distributed).

Where the maximum values of the characteristic-length ratios are too large to be easily displayed, the second greatest such ratio is shown, as a dashed line.

The first thing to notice is that the performance of the Fogel-Huang algorithm is very varied, ranging from producing a final ellipsoid with characteristic length 1.7750 times that of the corresponding minimum-volume ellipsoid (for a 5-dimensional data set with noise with a truncated

normal distribution with standard deviation $1/2\sqrt{3}$), to one with characteristic length 101.7132 times greater (for a 2-dimensional data set with uniformly distributed noise).

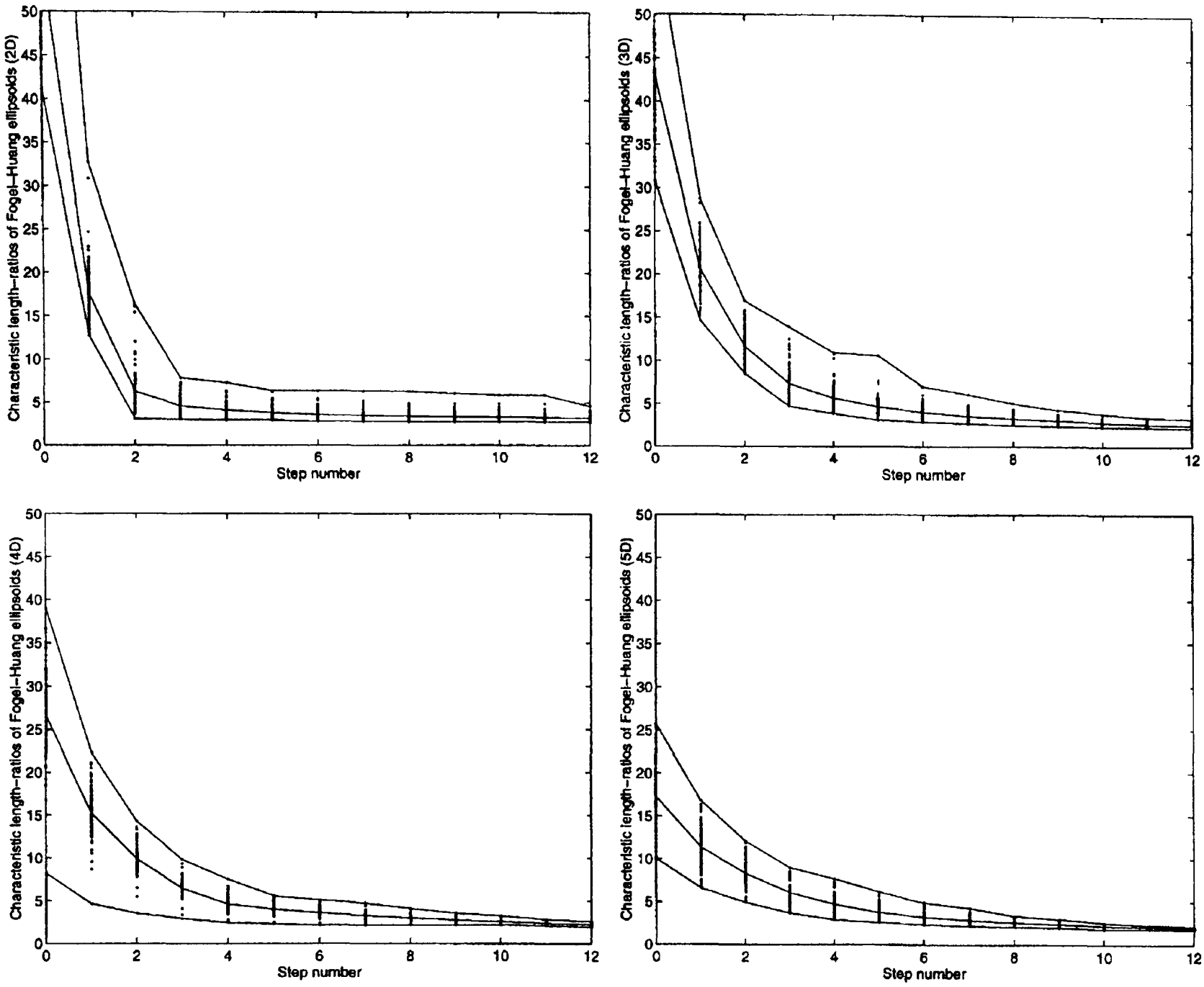


Figure 3.18: Characteristic lengths of Fogel-Huang ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

Secondly, the performance relative to that of the minimum-volume ellipsoid also appears to improve as the standard deviation of the noise decreases, that is, as the average angle between the normals to successive hyperplane pairs decreases. In particular, there is a great improvement in the worst cases.

Thirdly, a few very poor cases bring down the mean relative performance of the Fogel-Huang characteristic lengths with respect to the minimum-volume characteristic lengths, and there are far more of these poor cases when the noise has a greater spread. Table 3.2 gives information in support of this.

Finally, the relative performance also appears to improve as the dimension of the parameter

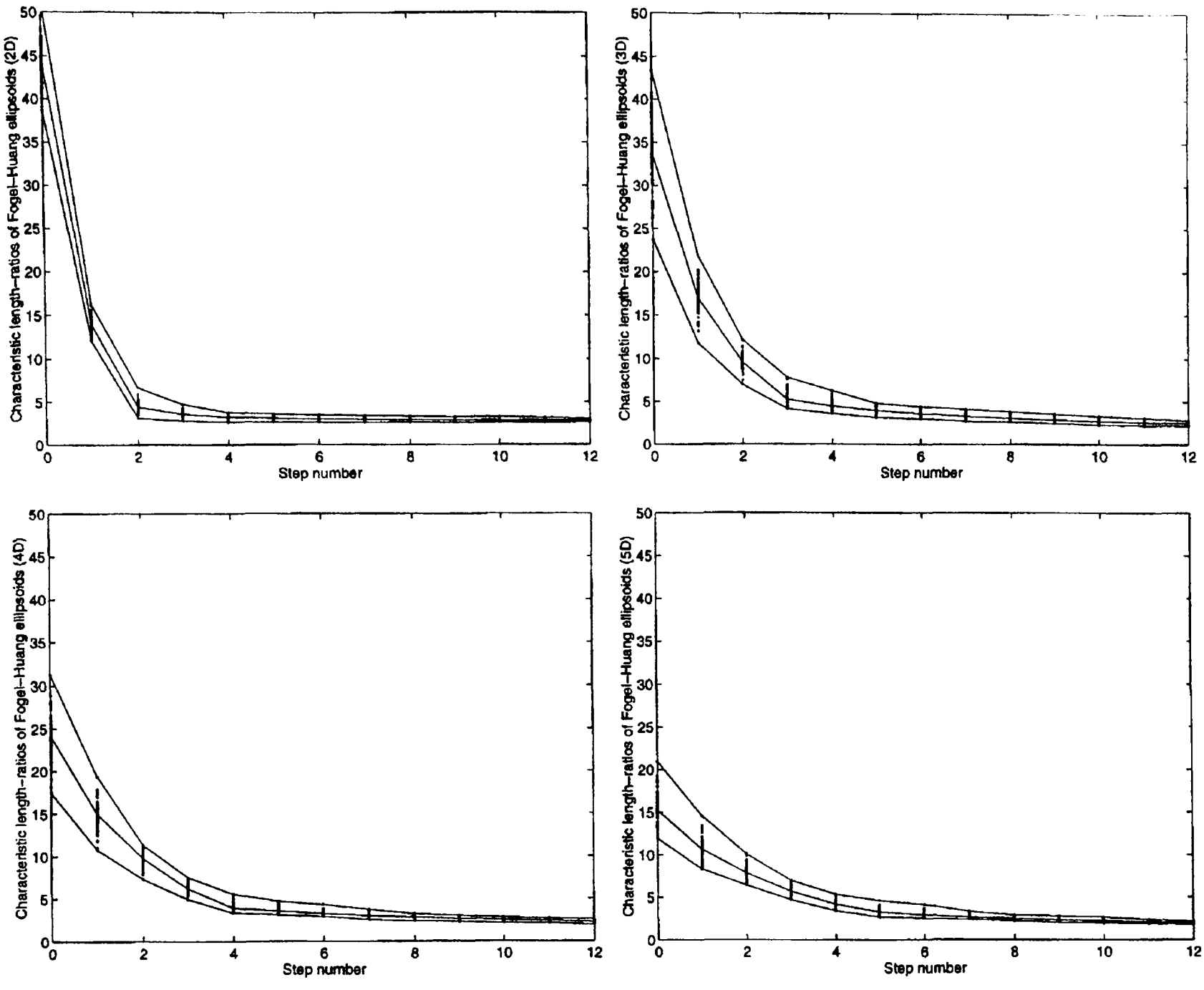


Figure 3.19: Characteristic lengths of Fogel-Huang ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

space increases.

Figures 3.20 to 3.22 are similar to Figures 3.14 to 3.16, but again the ellipsoids are the Fogel-Huang ellipsoids, and ratios (relative to the value for the final minimum-volume ellipsoid for the data set in question) replace absolute values.

The values shown are quite "wild", but this is to be expected, as neither the centre-parameter distance for the minimum-volume ellipsoid nor that for the Fogel-Huang ellipsoid is directly optimised, so the numerator in the ratio might "accidentally" be quite large and the denominator quite small, or *vice versa.*

Nevertheless, the performance of the centre of the Fogel-Huang ellipsoid as a point estimate of the true parameter compared to the performance of the centre of the minimum-volume ellipsoid improves both as the standard deviation of the noise affecting the data decreases, and as the

dimension of the parameter space increases, at least as far as concerns the mean of the results. Figures 3.23 to 3.25 show the characteristic-length ratios of the final Fogel-Huang ellipsoids (i.e., after the 12th step) plotted against the dimension of the parameter space. The improvement in performance with increasing dimension as measured by these ratios is thereby brought out. Figures 3.26 to 3.28 do the same thing for the improvement with increasing dimension of the centre of the Fogel-Huang ellipsoid as a point-estimate of the parameter, although this is only unambiguously evident for the mean performance.

To attempt to explain, at least for data affected by uniformly distributed noise, the improvement in the performance of the Fogel-Huang algorithm with increasing dimension, it will be necessary to make some assumptions:

1. at each step, the matrix Q_k can be replaced by $(\det Q_k)^{1/p} I_p = \overline{C_r^2} I_p$ without introducing enough error to invalidate the calculations;

2. the correlation between y_k and earlier y's for $k \geq 1$ caused by the dynamic nature of the model can be ignored. It is assumed that the y_k can be taken to be independently and uniformly distributed in $[-1+\bar{y}_k, 1+\bar{y}_k]$, where $\bar{y}_k = 0$, $k \leq 0$ and $\bar{y}_k = \hat{\theta}^p + \sum_{\ell=1}^{p-1} \hat{\theta}^\ell \bar{y}_{k-\ell} = (1 + \sum_{\ell=1}^{p-1} \bar{y}_{k-\ell})/\sqrt{p}$, where $\hat{\theta} = (1, \ldots, 1)/\sqrt{p}$ is now treated as known.

3. although the distribution of a sum of squares of variables uniformly distributed over intervals of equal length can be derived, this expression is complicated, and it is assumed that the sum of squares of $y_{k-1}, \ldots, y_{k-p+1}$ will be normally distributed with mean $\sum_{\ell=1}^{p-1} \overline{y_{k-\ell}^2}$ and standard deviation $\sqrt{\sum_{\ell=1}^{p-1} \sigma_{k-\ell}^2}$, where $\overline{y_k^2} = \bar{y}_k^2 + \frac{1}{3}$ is the mean, and $\sigma_k^2 = (4 + 60\bar{y}_k^2)/45$ the variance, of the distribution of the square of a variable uniformly distributed in $[-1 + \bar{y}_k, 1 + \bar{y}_k]$.

σ^2_{noise}	$1/\sqrt{3}$				$1/2\sqrt{3}$				$1/4\sqrt{3}$			
	Dimension				Dimension				Dimension			
Step	2	3	4	5	2	3	4	5	2	3	4	5
2	30	33	39	42	43	46	50	47	49	54	54	48
6	21	32	38	39	44	38	47	45	46	44	35	44
12	22	35	33	39	34	43	46	47	43	50	51	42

Table 3.2: Percentage of characteristic lengths exceeding the mean.

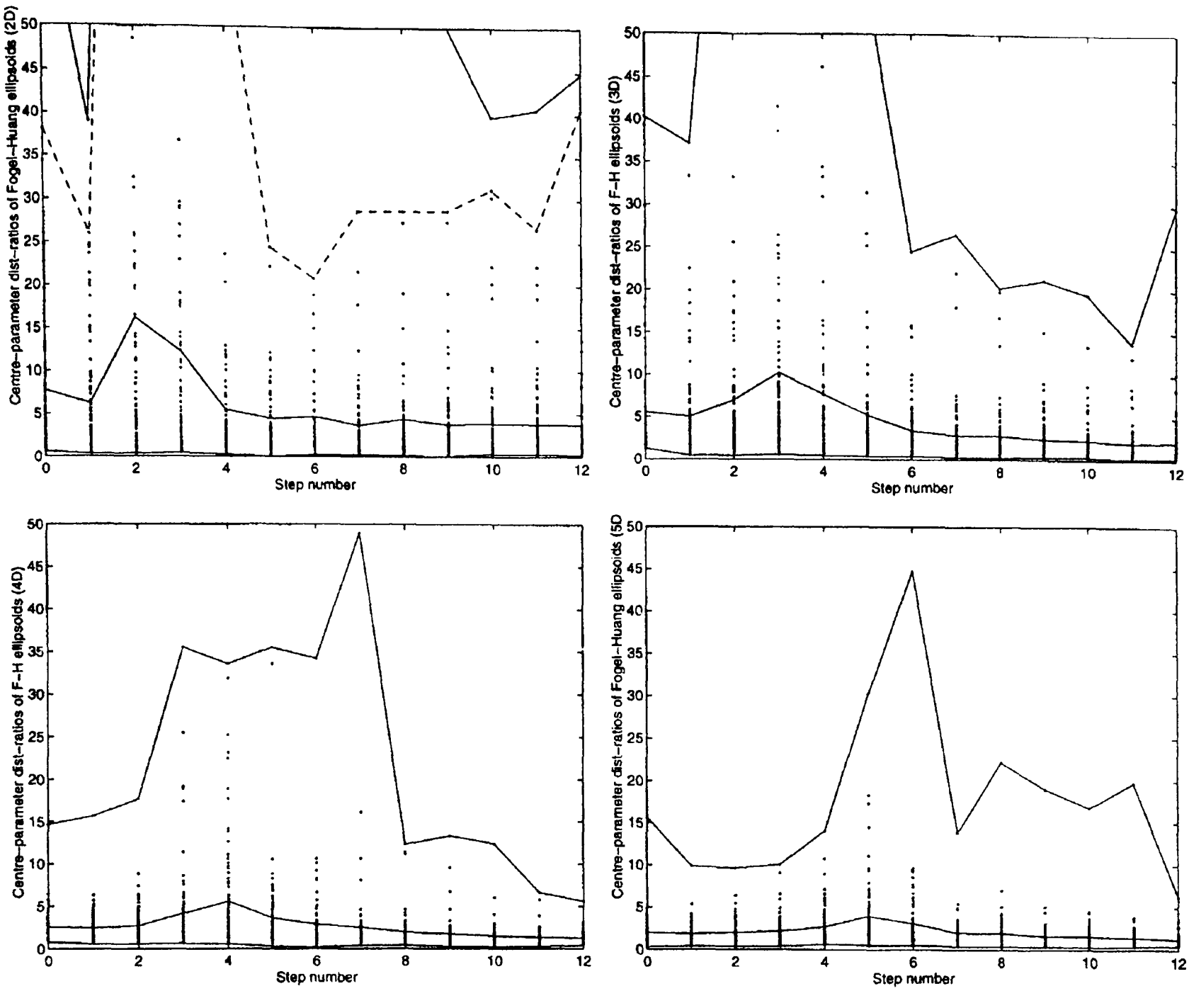


Figure 3.20: Ellipsoid centre to true parameter distance of Fogel-Huang ellipsoids (noise uniformly distributed).

The method is this: use the parameter vector to find the distribution of g_1, use the equations (3.40), (3.48), and (3.49) and (3.51), (3.54), and (3.58) to find $\bar{C}_r$ and $\overline{C_r^2}$ averaged over this distribution. $\bar{C}_r U_p^{1/p} (\det Q_0)^{1/2p}$ is then an estimate for the characteristic length after the first step and $\overline{C_r^2} (\det Q_0)^{1/p} I_p$ is an estimate for Q_1 (U_p being the volume of the unit sphere in p dimensions).

The estimate probability density function for g_1 is that of a random variable equal to $400 U_p^{2/p}$ (the initial square characteristic length) times the sum of an ordinary variable with value 1 (i.e., u_1) and $p-1$ normally distributed random variables (i.e., $y_0, \ldots, y_{-p+2}$), each with mean $\frac{1}{3}$ and variance $4/45$, so g_1 has a Gaussian estimate probability density function with mean $400(1+\frac{p-1}{3})$ and variance $160000 \times 4(p-1)/45$. As these means are many times the corresponding standard deviations away from zero, the effect of the correction that is made to take account of the

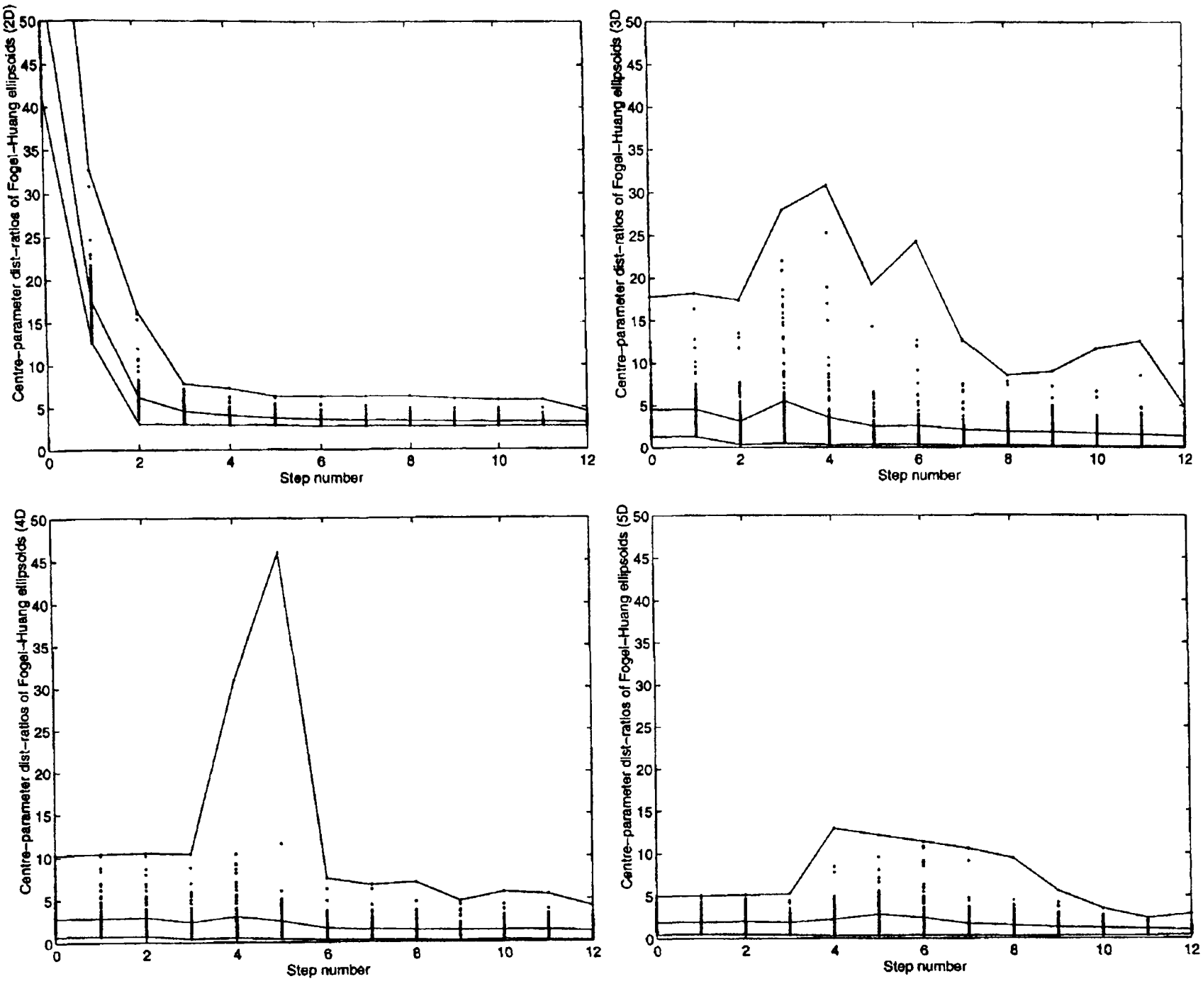


Figure 3.21: Ellipsoid centre to true parameter distance of Fogel-Huang ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

fact that g_1 cannot be negative (multiplying by a constant to ensure that the integral of the left-truncated normal distribution is unity) is negligible, but this will not be the case later on. The probability density functions for various values of p are shown in Figure 3.29.

In Figures 3.30 and 3.31 we redisplay $\bar{C}_r$ and $\overline{C_r^2}$ from Figures 3.7 and 3.9, but with $P(g_1)$, the cumulative probability function of g_1 along the horizontal axis. We (numerically) calculate the areas under these curves to give estimates of $\bar{C}_r$ and $\overline{C_r^2}$ averaged over g_1. The initial characteristic lengths are multiplied by the averaged $\bar{C}_r$ to give estimates of the characteristic lengths after the first application of the Fogel-Huang algorithm, and the squares of the initial characteristic lengths are then multiplied by the estimate for $\overline{C_r^2}$ to give estimates for the squares of the characteristic lengths, also after the first step. The estimate for each p, Q_1, is then the identity matrix in p dimensions, I_p, multiplied by the appropriate estimate for the squared

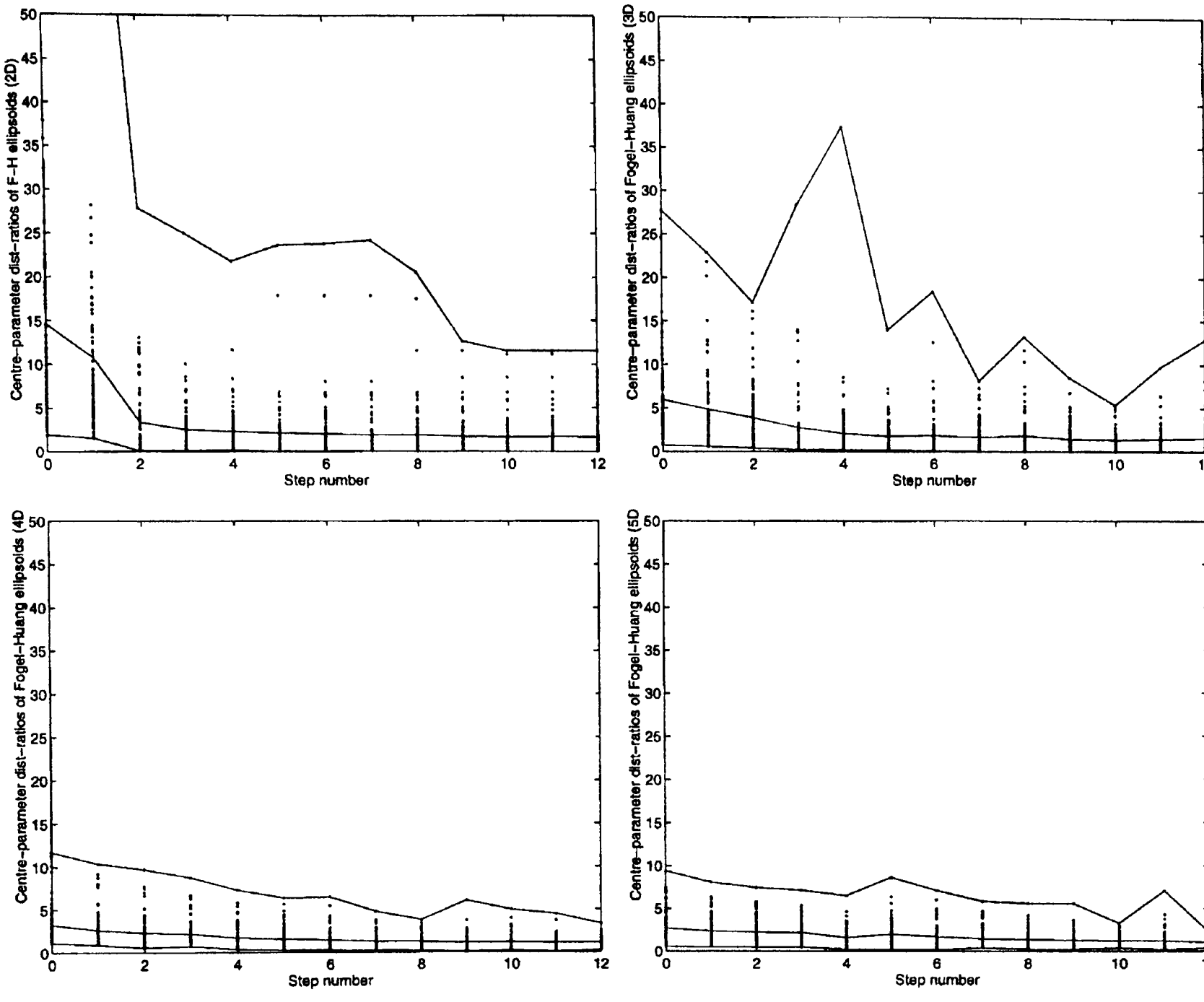


Figure 3.22: Ellipsoid centre to true parameter distance of Fogel-Huang ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

characteristic length, divided by $U_p^{2/p}$.

The estimate for the distribution of g_2 is that of the sum of the following variables, multiplied by the estimate for the square of the characteristic length: one non-probabilistic variable, with value 1; $p-2$ random variables with mean $\frac{1}{3}$ and variance $\frac{4}{45}$ (i.e., $y_0, \ldots, y_{-p+3}$); and one random variable with mean $\frac{1}{3} + \frac{1}{p}$ and variance $\frac{4}{45} + \frac{4}{15p}$ (i.e., y_1).

The probability density functions for g_2 are shown in Figure 3.32, and $\bar{C}_r$ and $\overline{C_r^2}$ are plotted against the corresponding cumulative probability functions in Figures 3.33 and 3.34.

Carrying on in this fashion leads to Figures 3.35 to 3.37 referring to g_4 and to Figures 3.38 to 3.40 referring to g_{11}, and Figure 3.41 showing the estimated characteristic lengths derived from this approach.

From these diagrams an idea of what is going on can be derived. As k increases, g_k decreases,

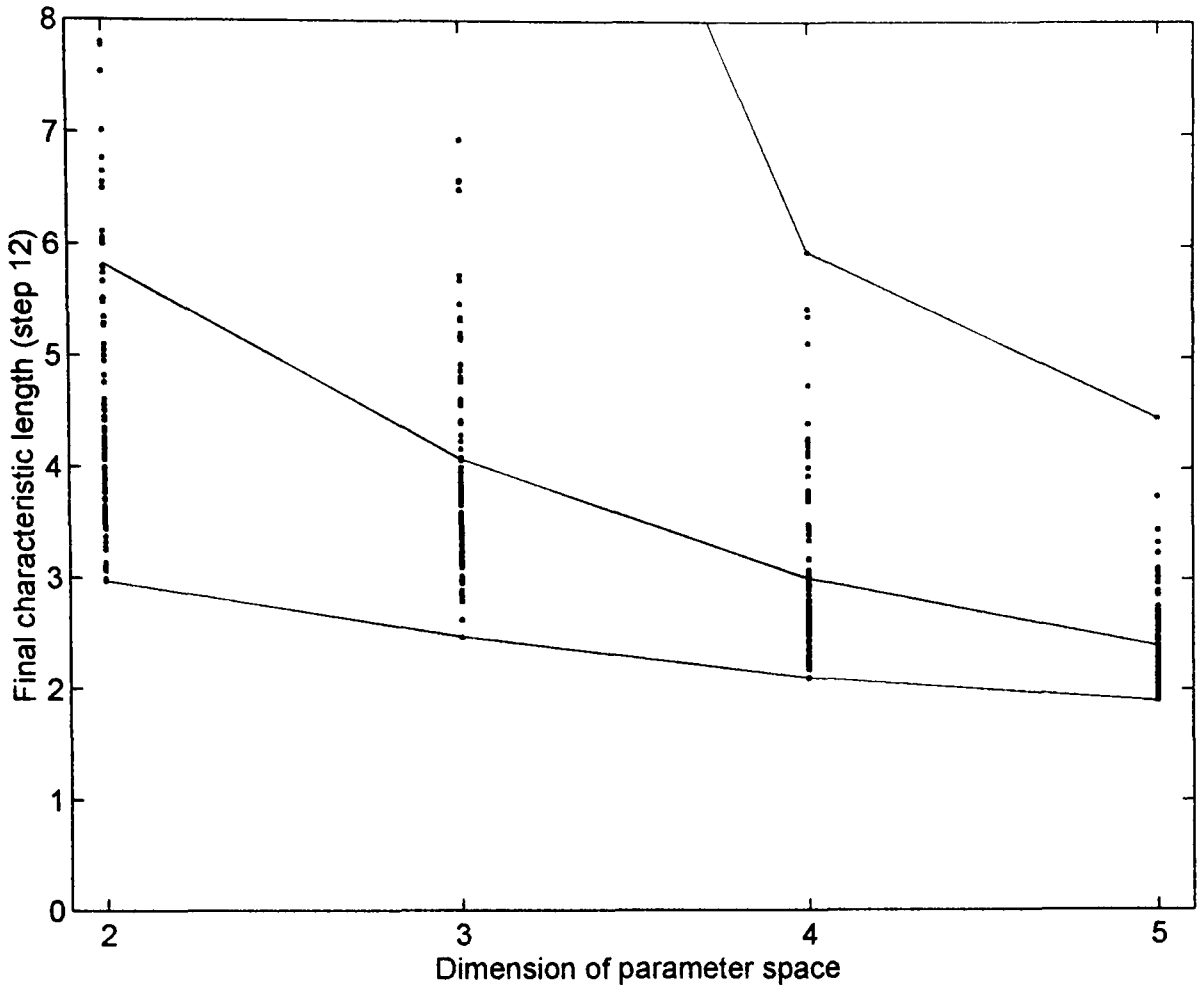


Figure 3.23: Final characteristic lengths for the Fogel-Huang algorithm (uniformly distributed noise).

but, for higher dimensional parameter spaces, this decrease is not so great, as the sum of the y's, as well as the y's themselves, increases more rapidly for higher p. It can be seen from Figures 3.29, 3.32, 3.35 and 3.38 that the median of the estimate probability function increases for higher p's relative to that for lower p's, and $\bar{C}_r$ and $\overline{C_r^2}$ averaged over g_k will be close to their values at these medians. The result of this is that although for fixed g_k, $\bar{C}_r$ and $\overline{C_r^2}$ decrease as p increases, it can be conjectured that $\bar{C}_r$ and $\overline{C_r^2}$ averaged over g_k do not decrease as p increases if k is sufficiently large. By the fourth step, the value of $\bar{C}_r$ at the median of g_4 for $p = 2$ exceeds its value at the median for $p = 3$, and the value of $\overline{C_r^2}$ at the median of g_4 for $p = 2$ exceeds its value at the median for $p = 3$ and 4, and by the 12th step, the values of $\bar{C}_r$ at the median values of g_{11} corresponding to $p = 5, 4, 3, 10, 2$ are in increasing order. The order for $\overline{C_r^2}$ is the same.

The estimate characteristic lengths displayed in Figure 3.41 are proportional to the products of the $\overline{C_r^2}$'s for the corresponding p up to the step number labelling the horizontal axis. It will be observed that by the 9th step the estimate characteristic length for $p = 2$ is greater than that for $p = 3$, by the 11th step it exceeds the estimate characteristic length for $p = 4$, and by the 13th step that for $p = 5$ has been overtaken. Also, although it is not visible to the resolution

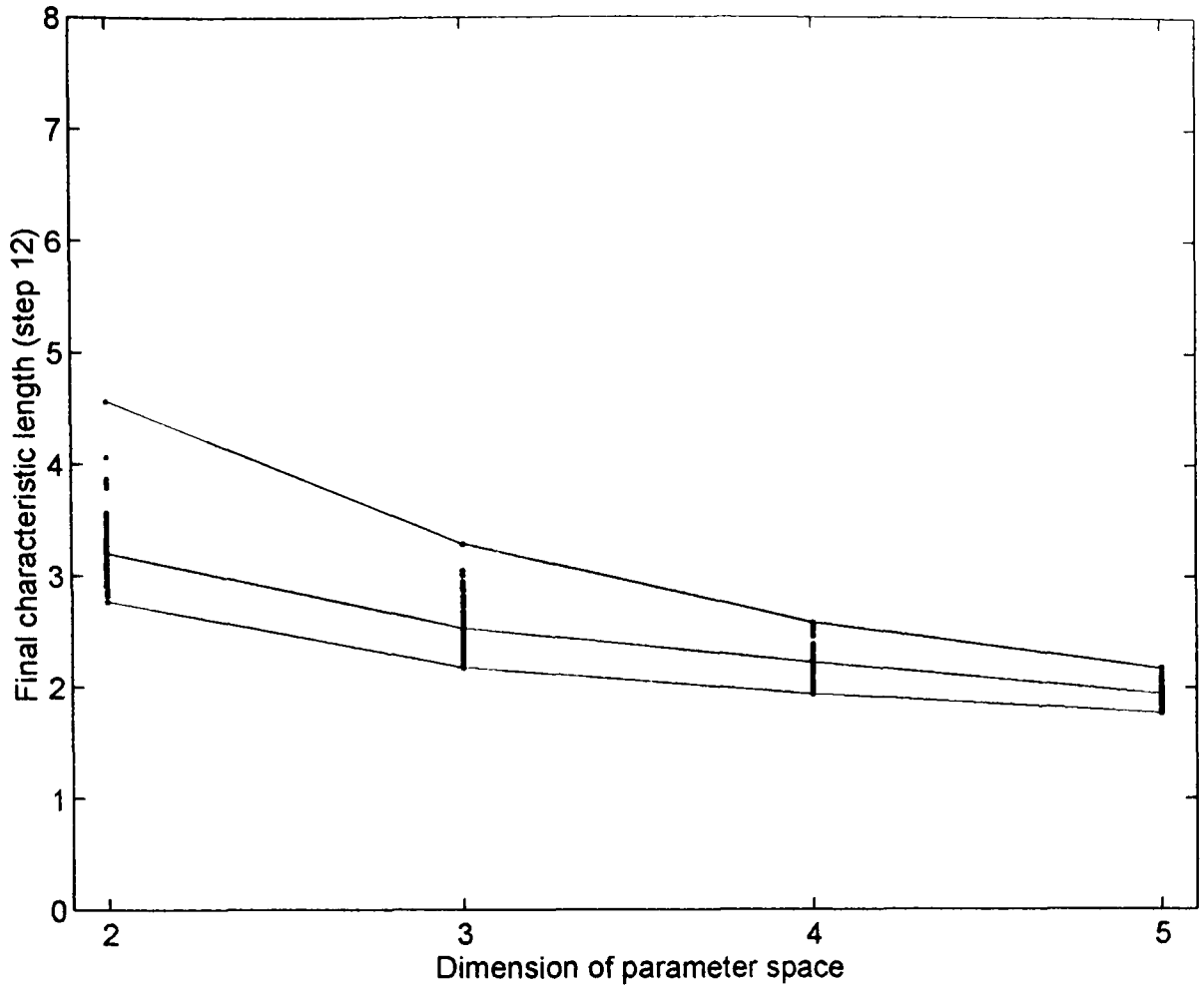


Figure 3.24: Final characteristic lengths for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

of the Figure, the estimate for $p = 3$ at the 15th step is greater than that for $p = 4$. Thus, the estimate displays the same trend as the empirical results, but this trend is gentler in the estimate.

Figures 3.42 and 3.43 show just how bad the Fogel-Huang ellipsoids can be. The BLJ (minimum-volume) ellipsoid is dwarfed by the Fogel-Huang ellipsoids after 6, 10 and 11 steps (corresponding to the introduction of the active hyperplanes — those bounding the feasible parameter set) and the 12th step produces no change in the ellipsoid. The ratio between the characteristic lengths of the final Fogel-Huang ellipsoid and the BLJ ellipsoid is 101.7132, whereas the ratio between the characteristic lengths of the BLJ ellipsoid and the polytope is 1.5551.

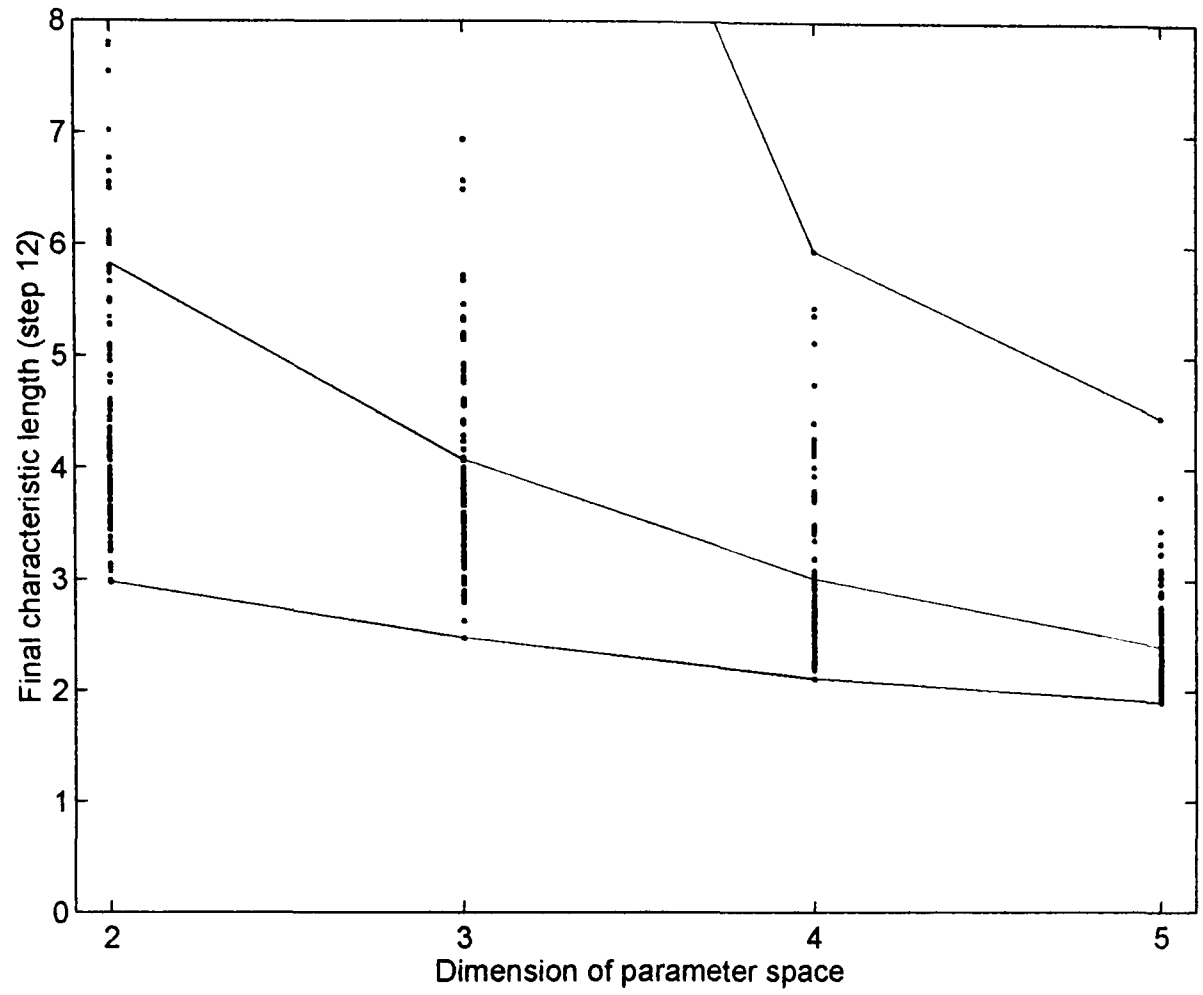


Figure 3.25: Final characteristic lengths for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

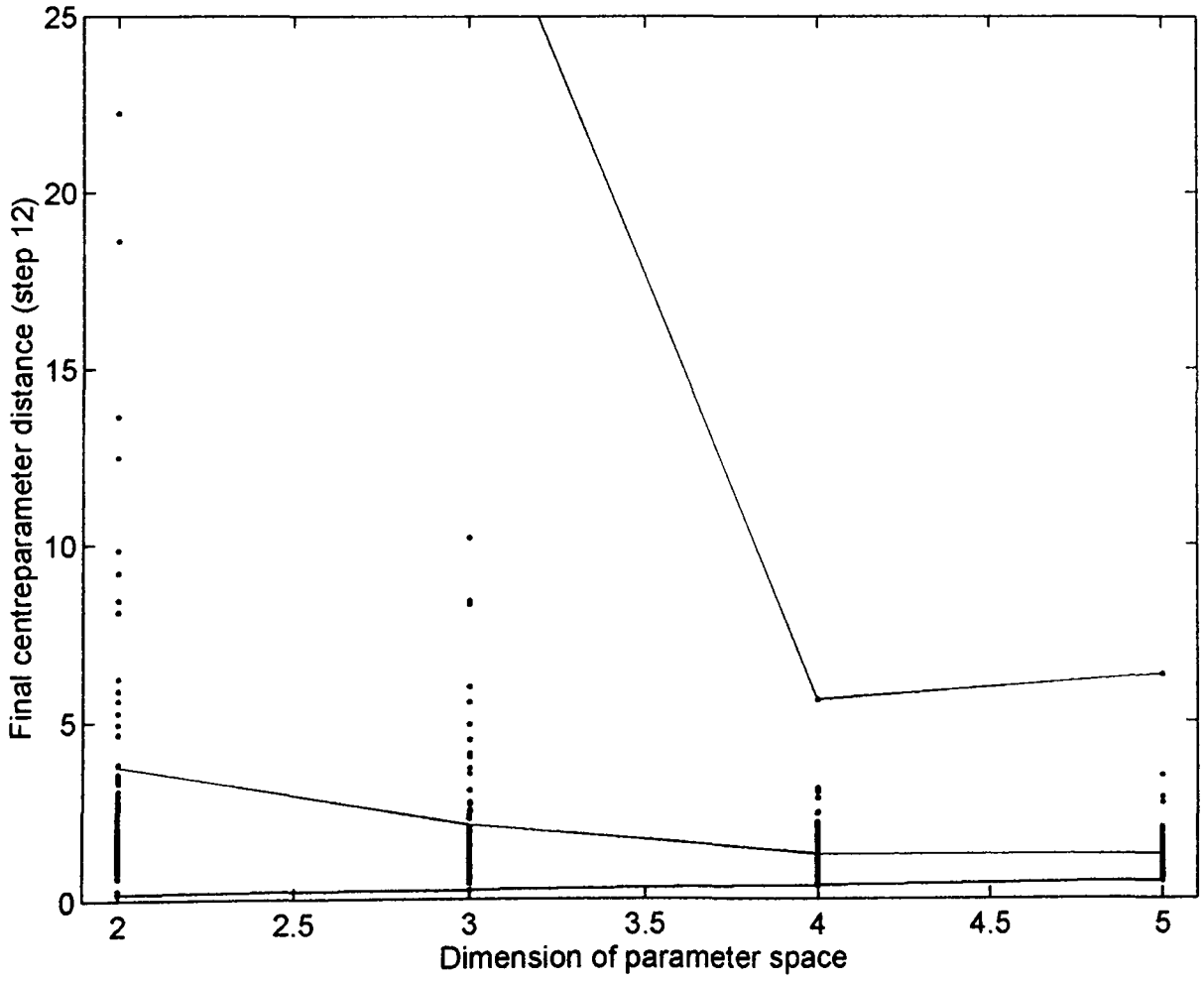


Figure 3.26: Final centre-parameter distance for the Fogel-Huang algorithm (uniformly distributed noise).

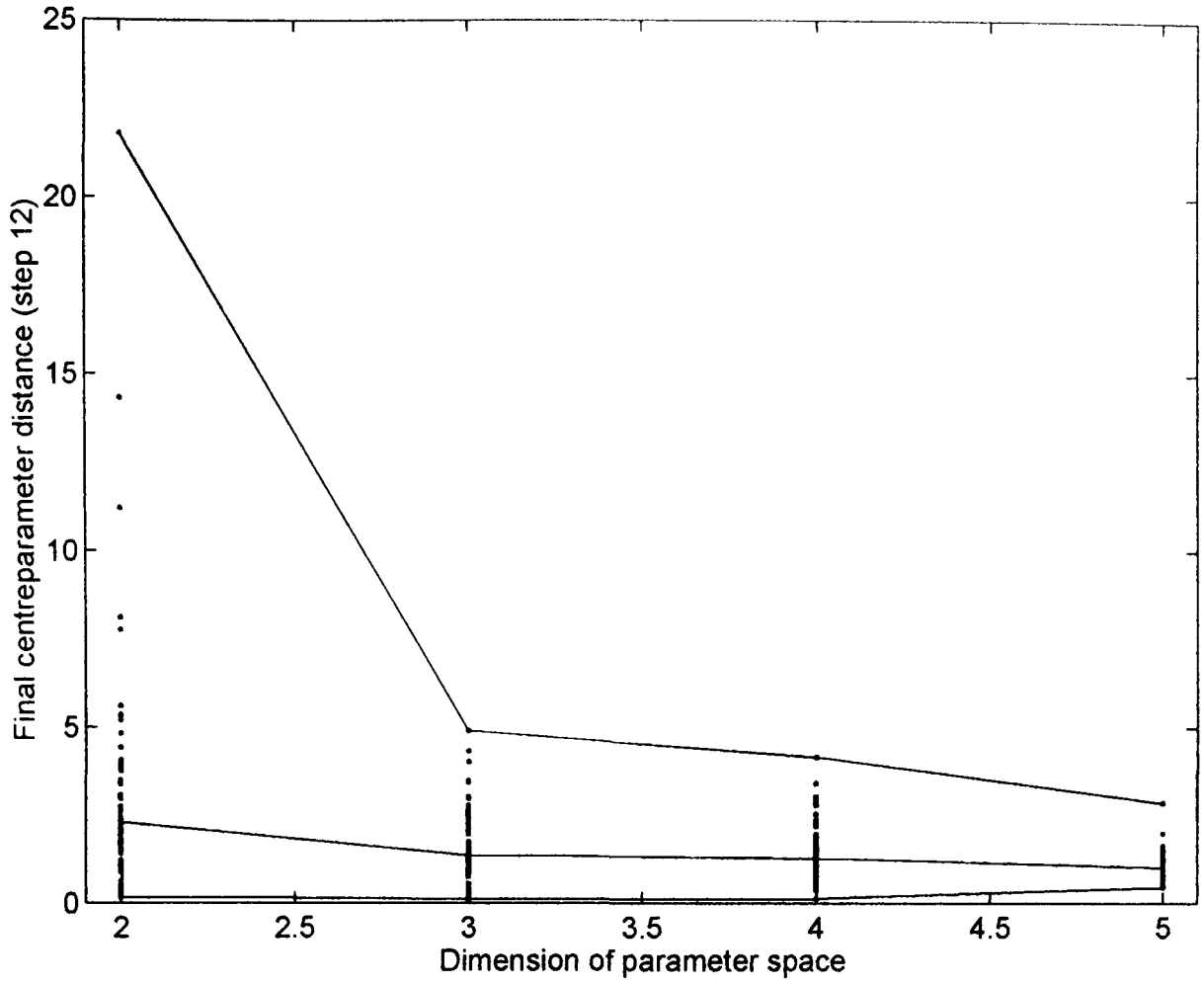


Figure 3.27: Final centre-parameter distance for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

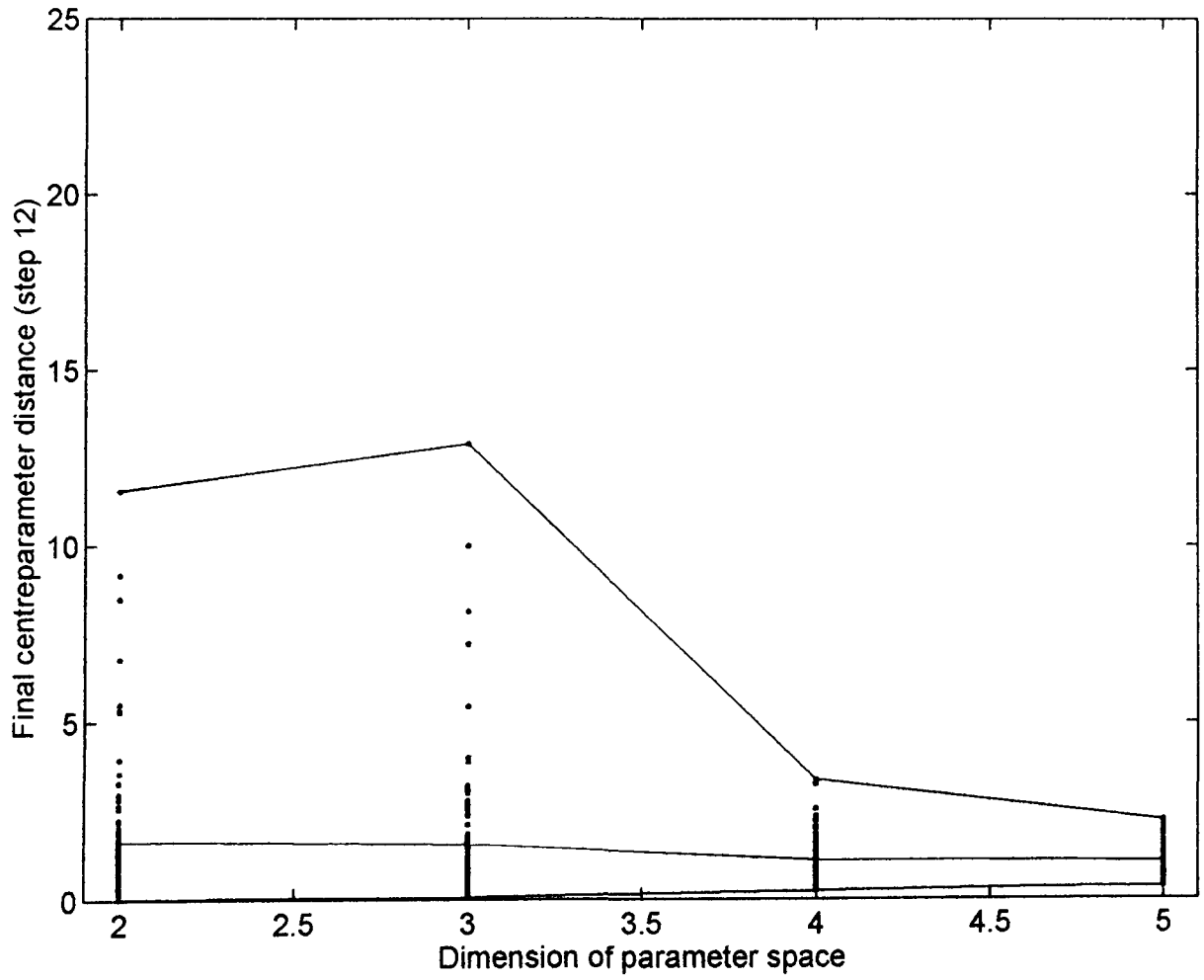


Figure 3.28: Final centre-parameter distance for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

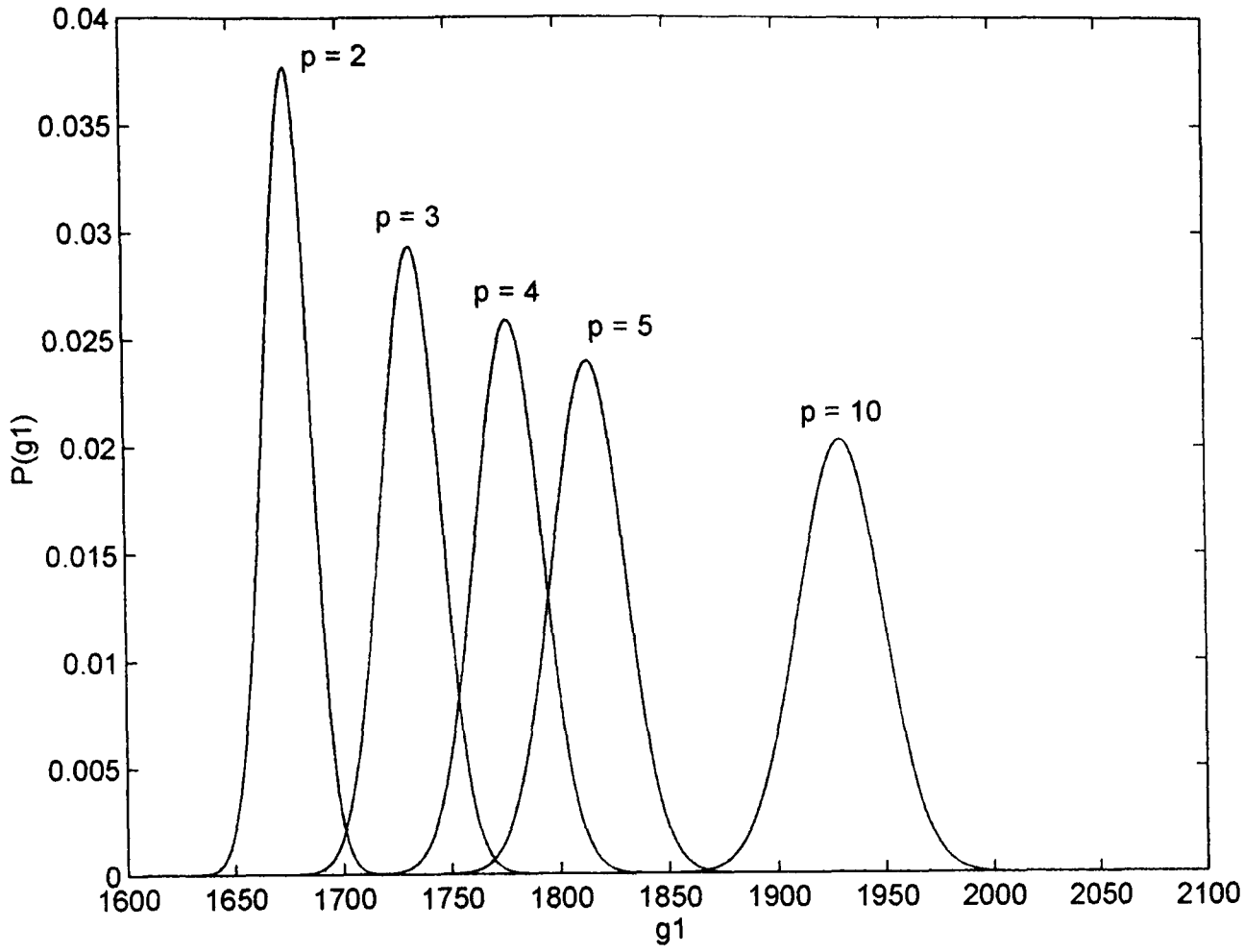


Figure 3.29: Estimate probability density function for g_1.

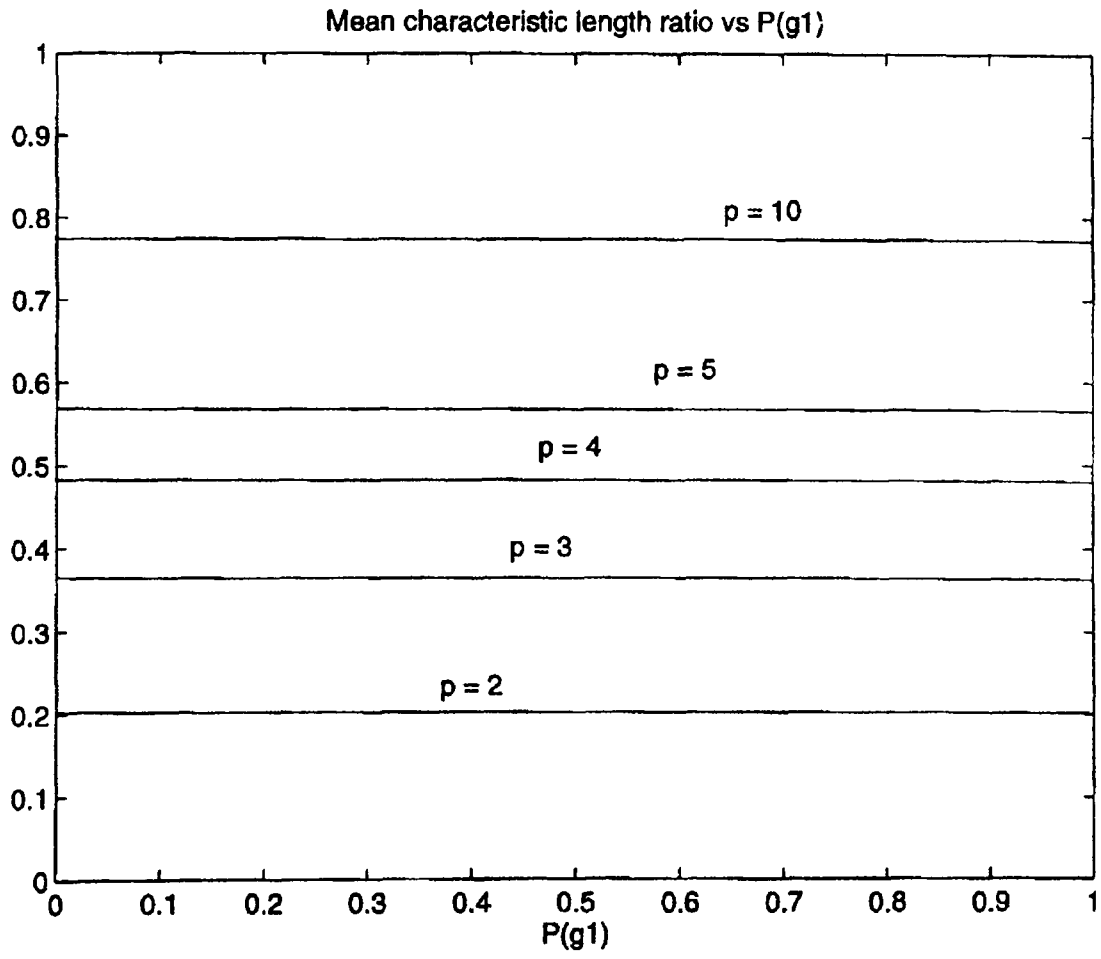


Figure 3.30: $\bar{C}_r$ against the (estimated) cumulative probability function for g_1.

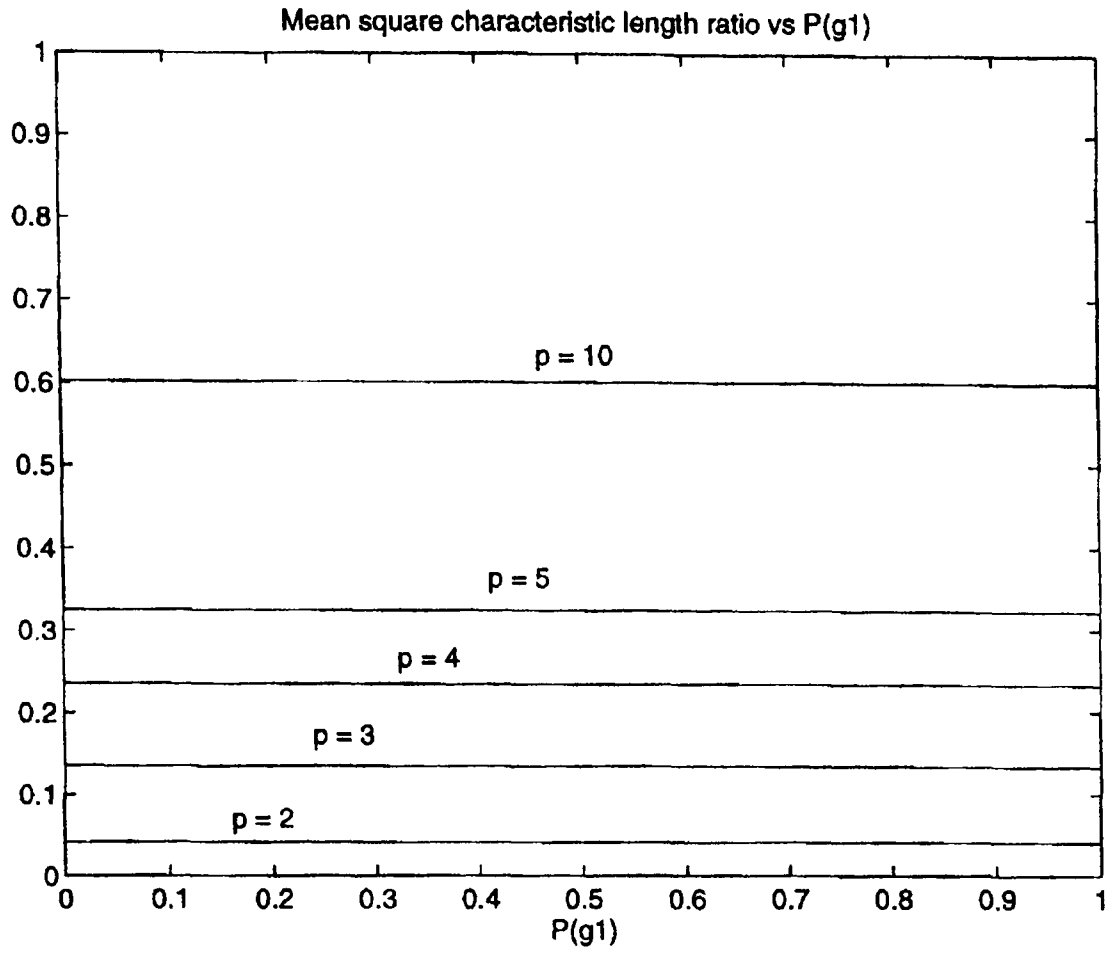


Figure 3.31: $\overline{C_r^2}$ against the (estimated) cumulative probability function for g_1.

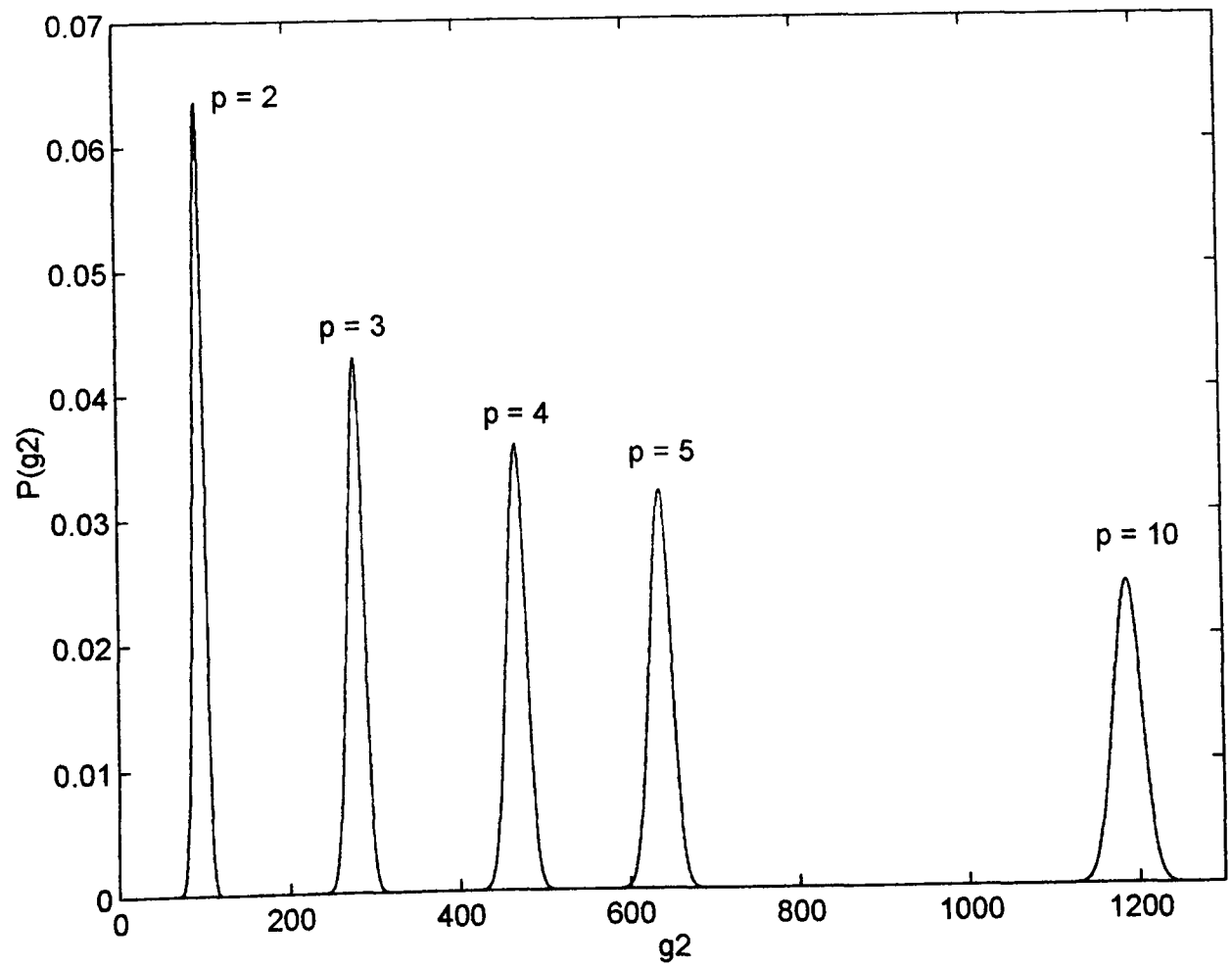


Figure 3.32: Estimate probability density function for g_2.

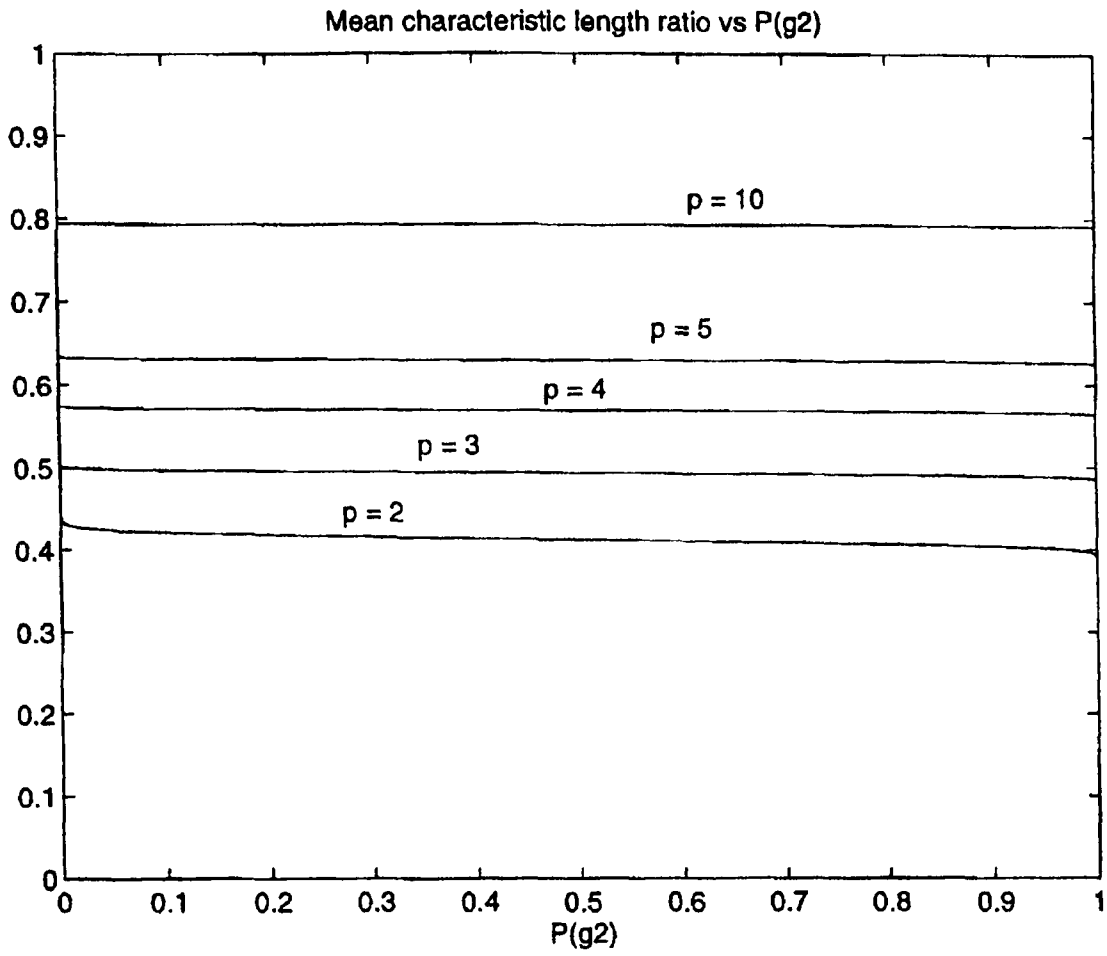


Figure 3.33: $\bar{C}_r$ against the (estimated) cumulative probability function for g_2.

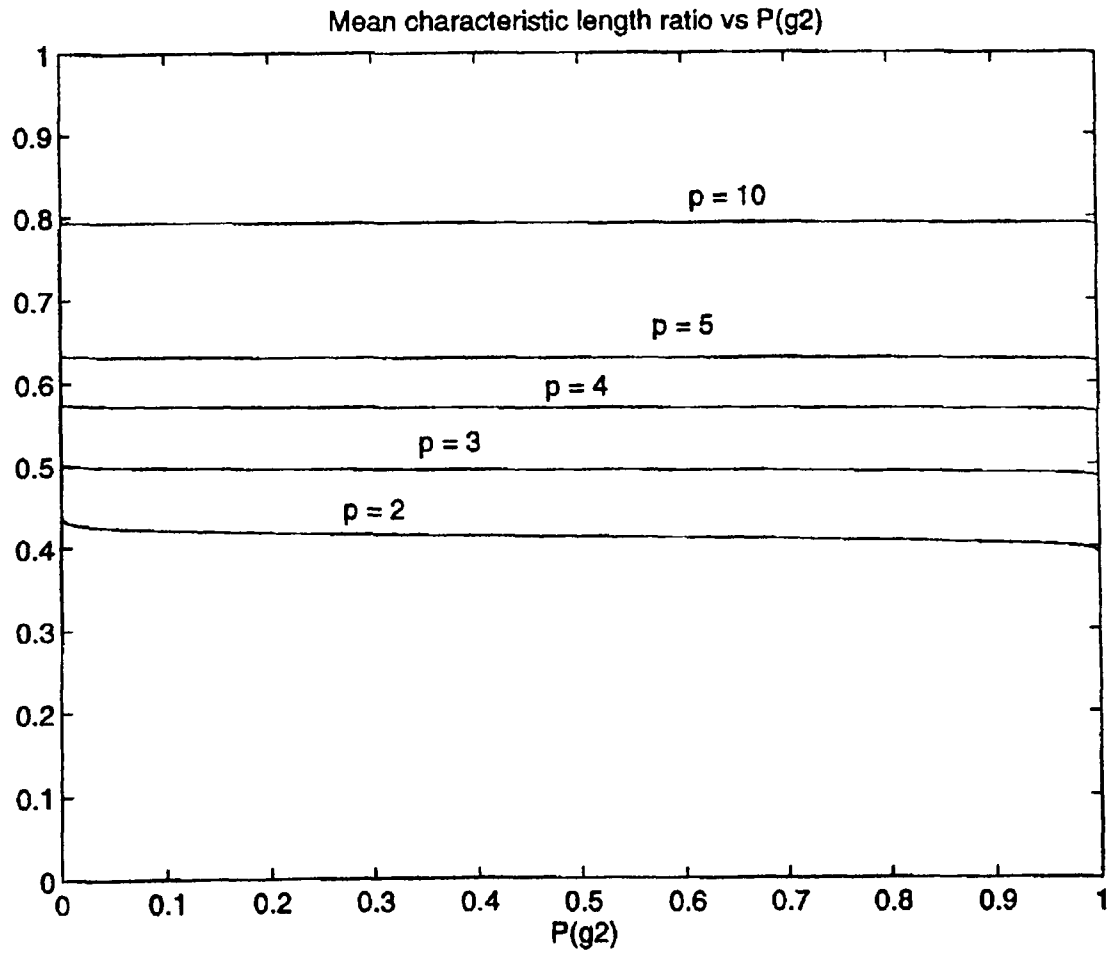


Figure 3.34: $\overline{C_r^2}$ against the (estimated) cumulative probability function for g_2.

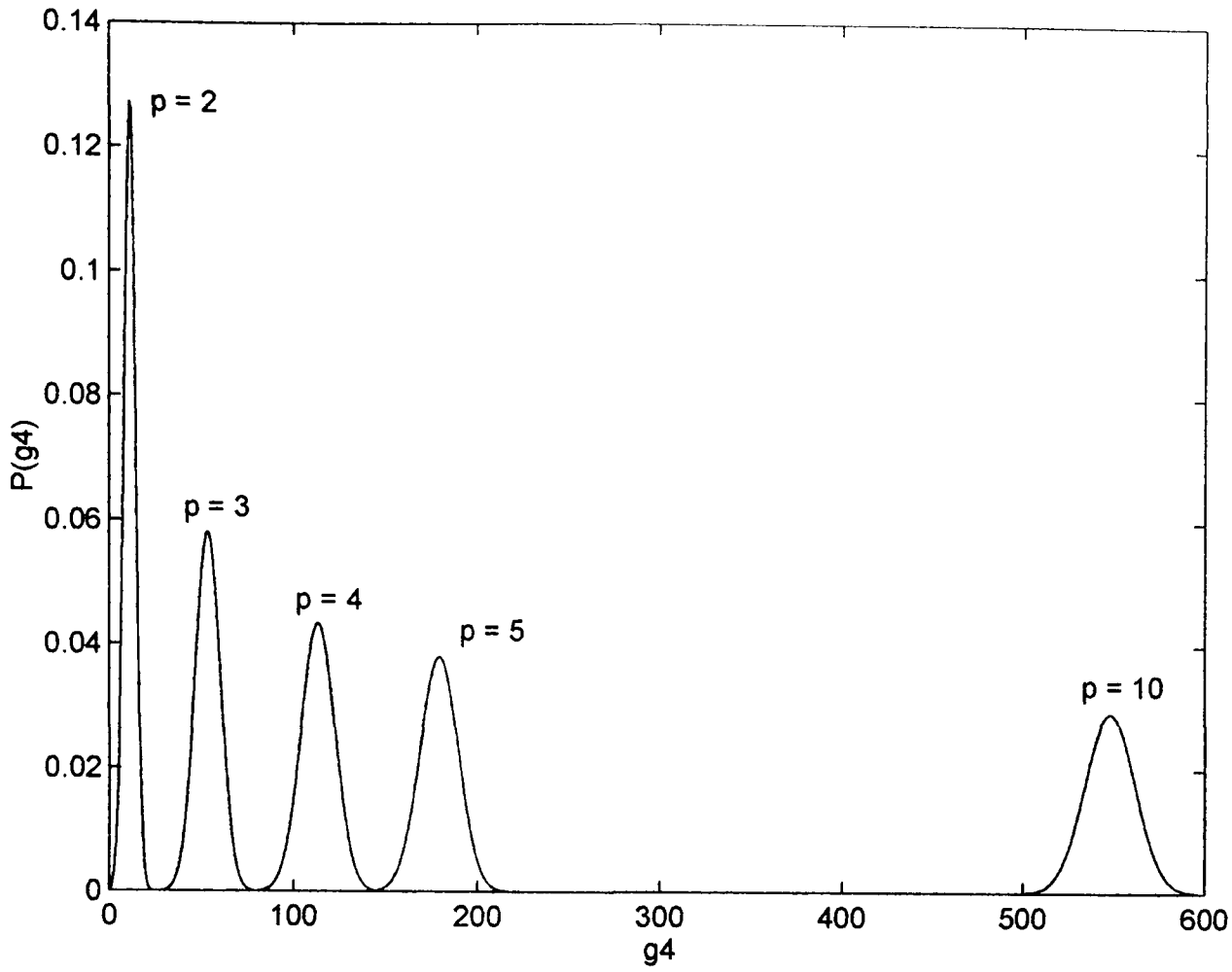


Figure 3.35: Estimate probability density function for g_4.

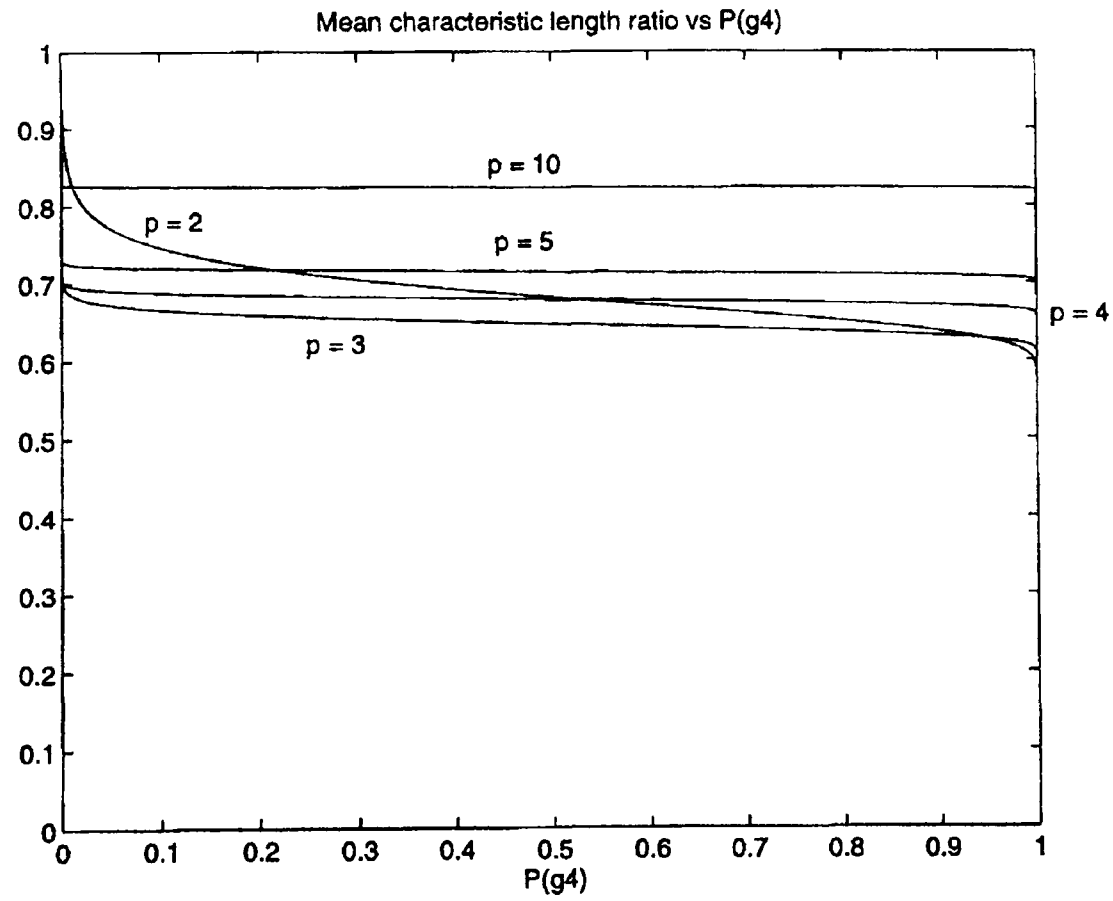


Figure 3.36: $\bar{C}_r$ against the (estimated) cumulative probability function for g_4.

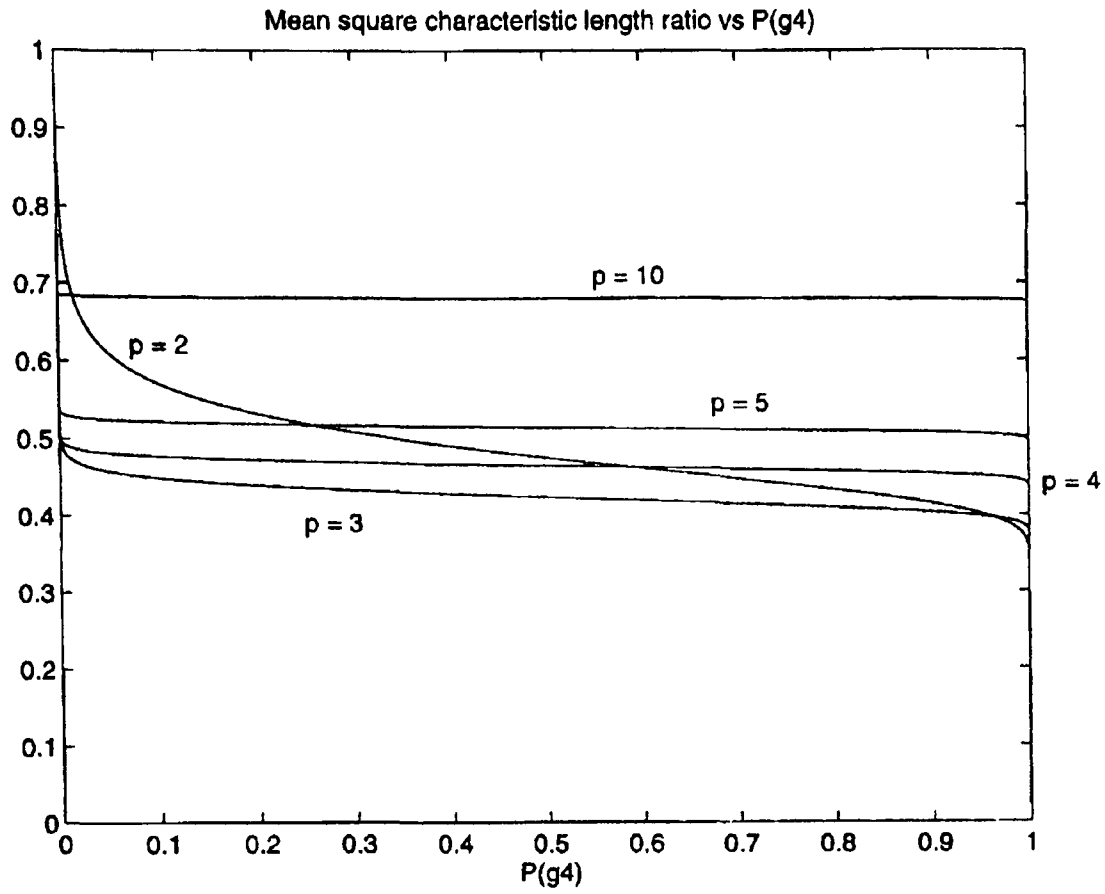


Figure 3.37: $\overline{C_r^2}$ against the (estimated) cumulative probability function for g_4.

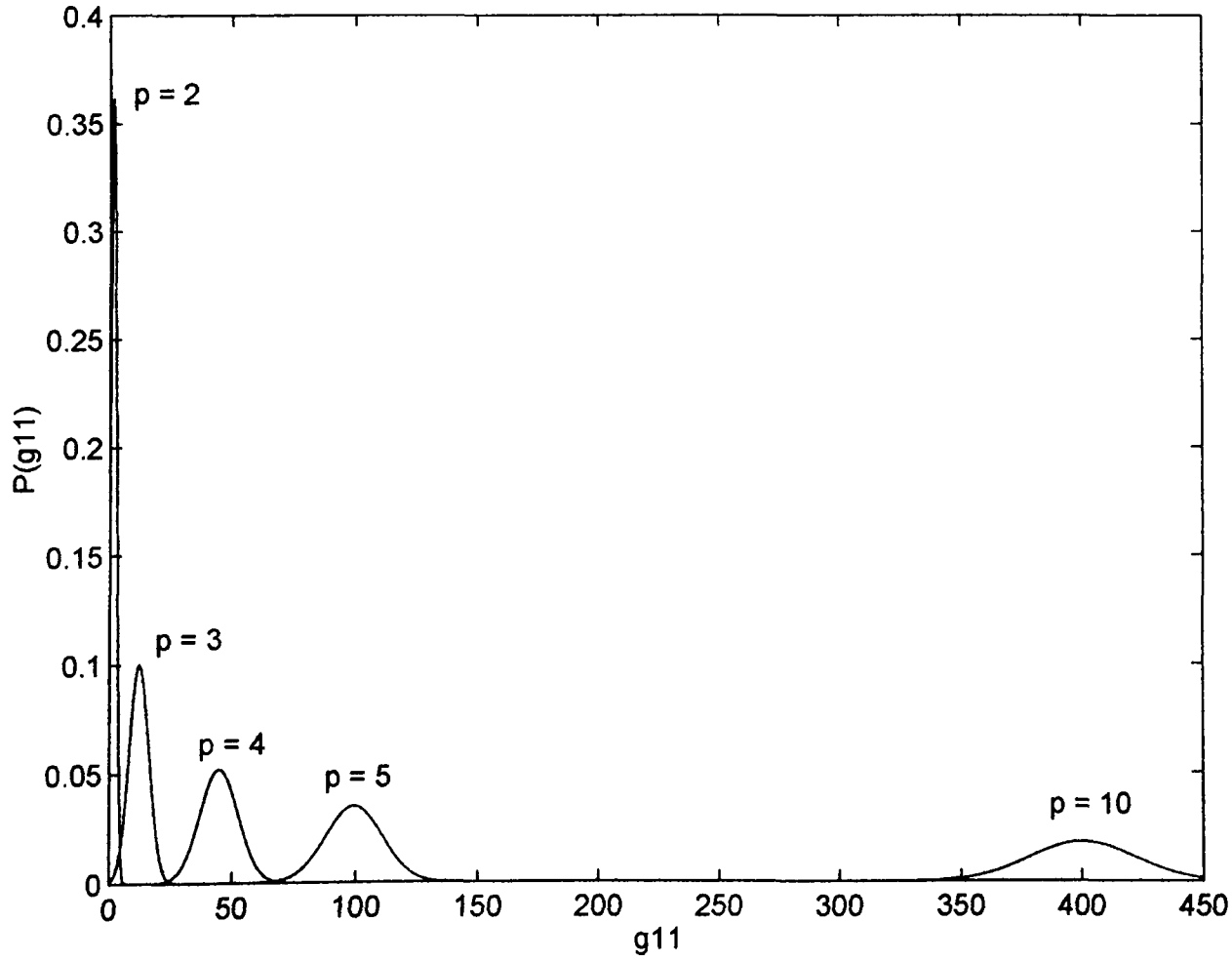


Figure 3.38: Estimate probability density function for g_{11}.

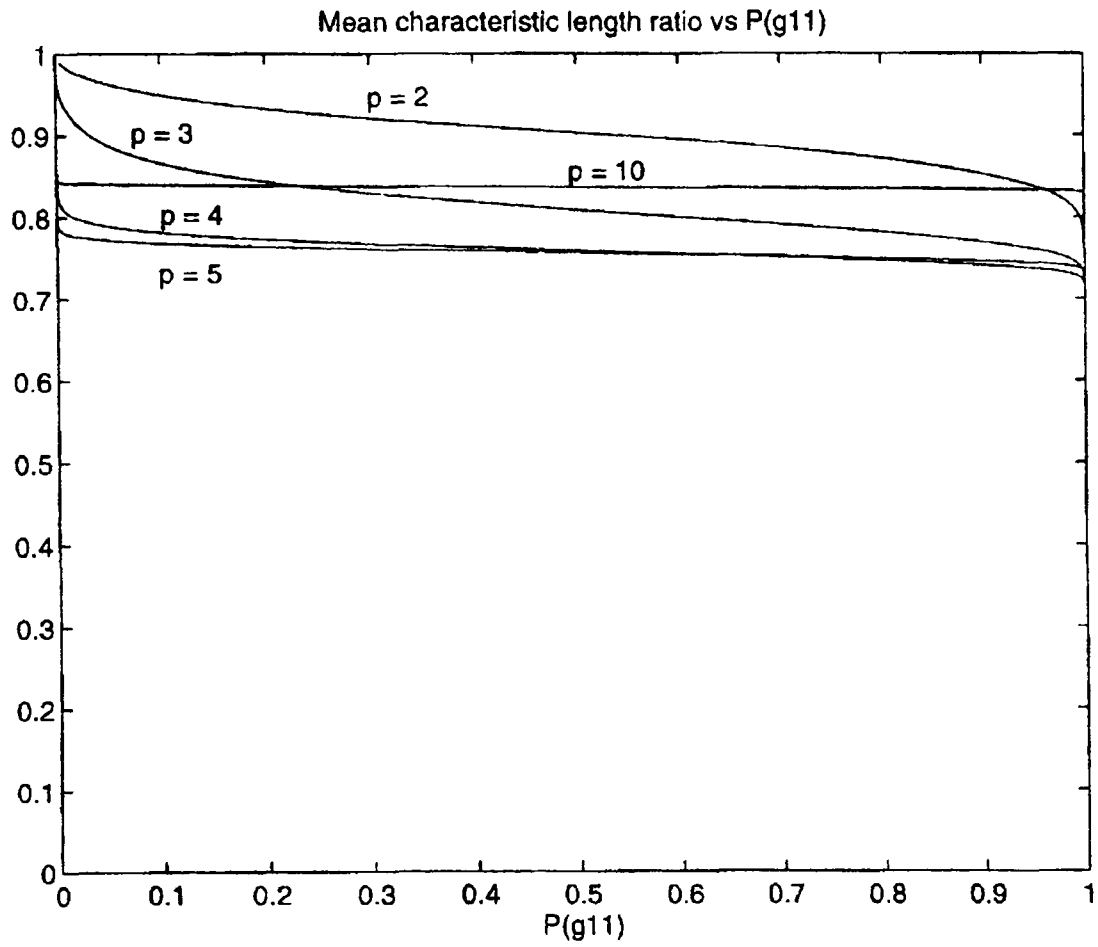


Figure 3.39: $\bar{C}_r$ against the (estimated) cumulative probability function for g_{11}.

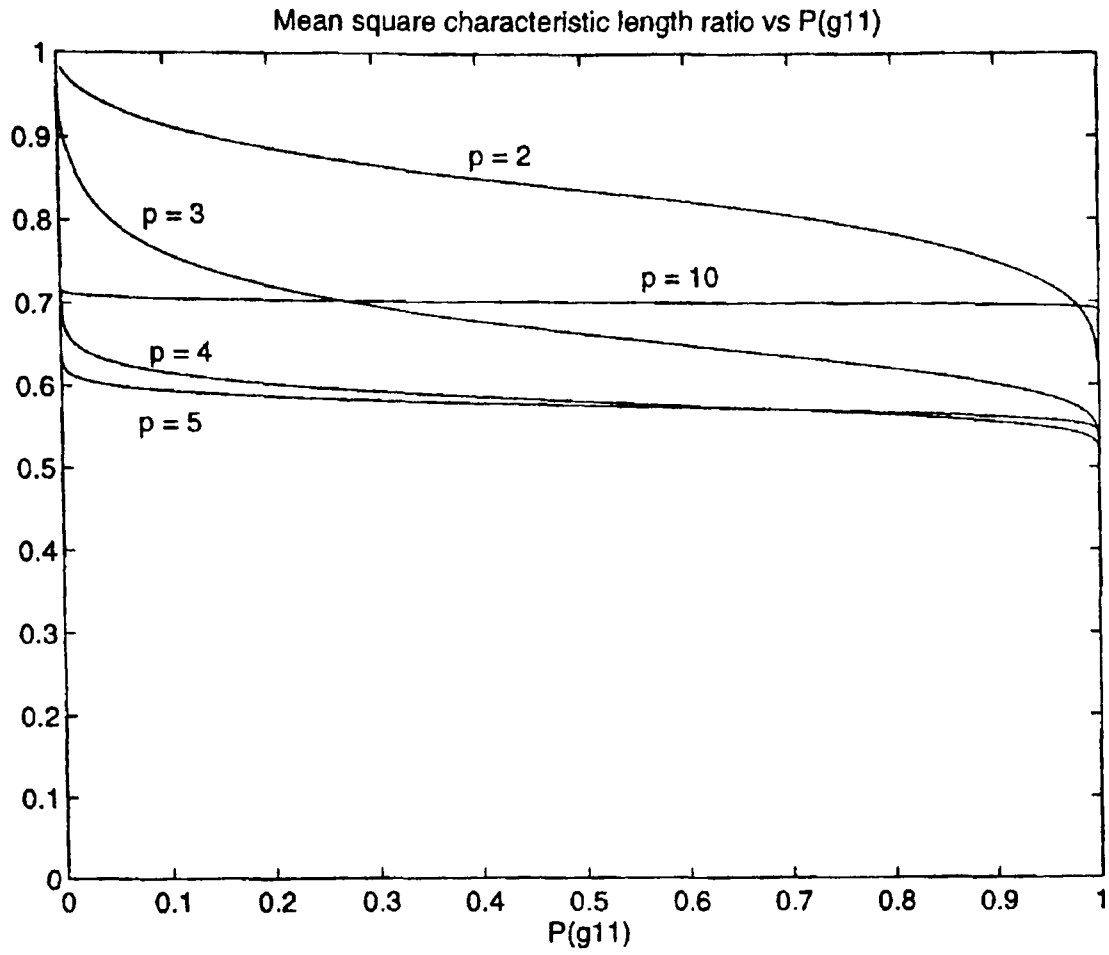


Figure 3.40: $\overline{C_r^2}$ against the (estimated) cumulative probability function for g_{11}.

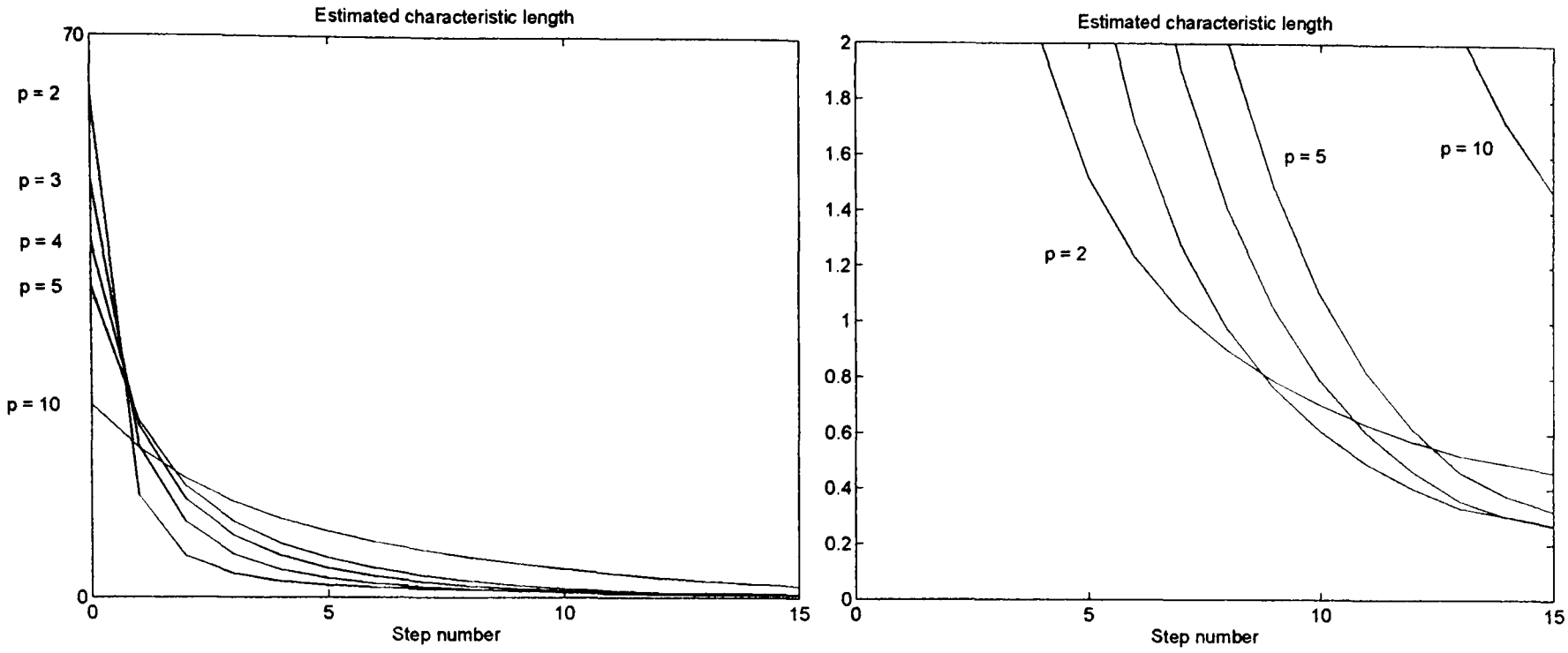


Figure 3.41: Estimated characteristic lengths (and detail).

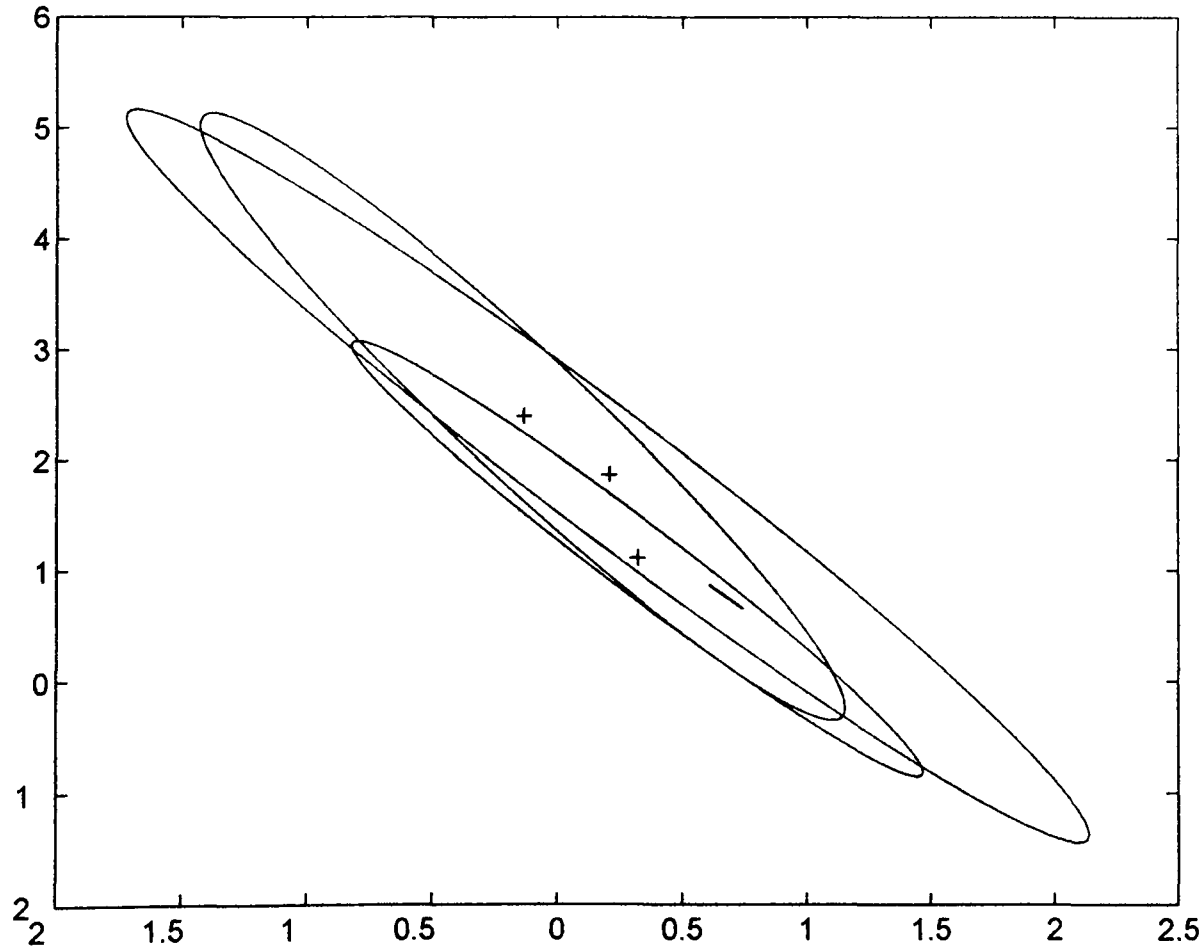


Figure 3.42: Fogel-Huang ellipsoids (for steps 6, 10 and 11) and BLJ ellipsoid (which appears as a line to the resolution of this Figure) for the two-dimensional data set leading to the greatest ratio between the volumes of the final Fogel-Huang ellipsoid and the BLJ ellipsoid.

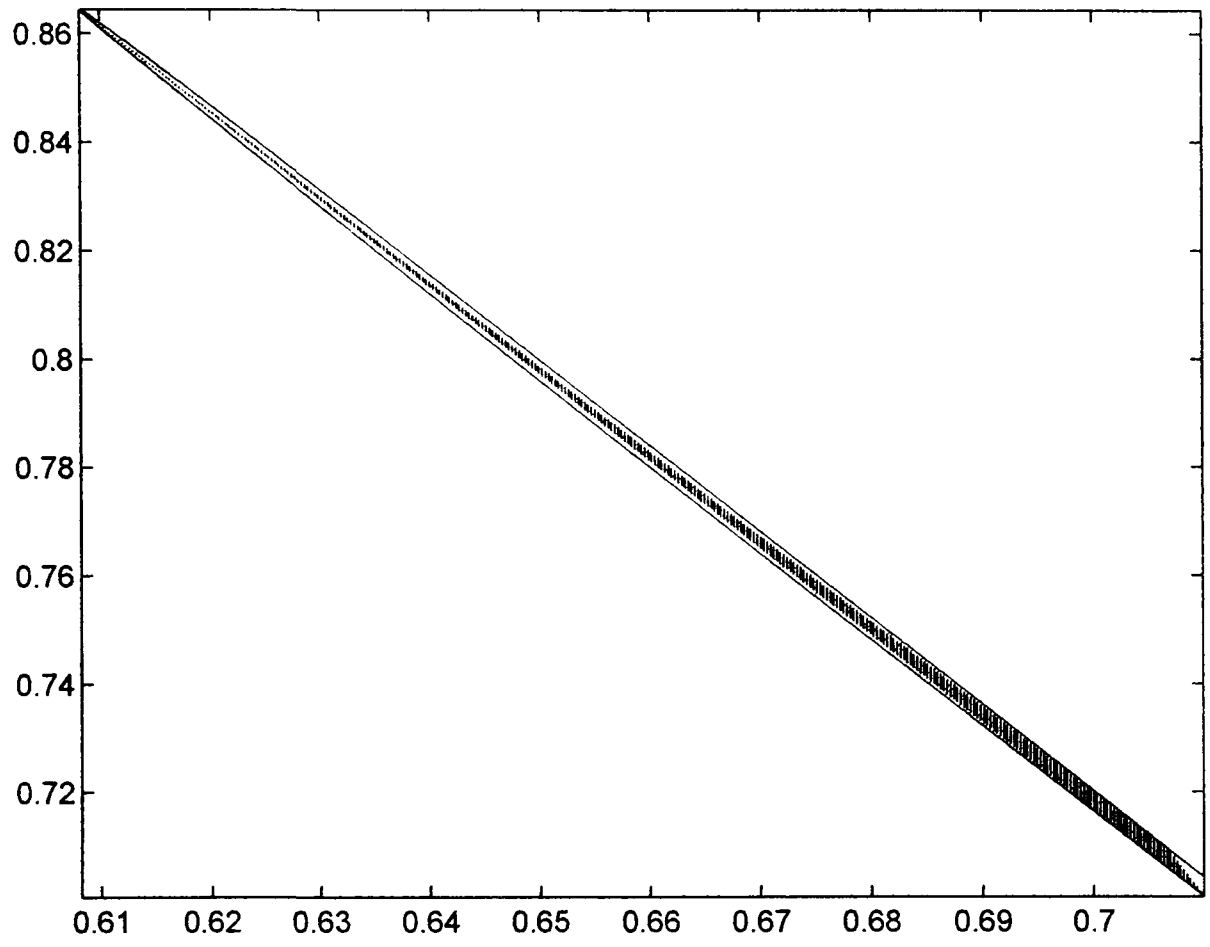


Figure 3.43: BLJ ellipsoid and polytope (shaded) for the two-dimensional data set leading to the greatest ratio between the volumes of the final Fogel-Huang ellipsoid and the BLJ ellipsoid.

3.3.3 With Data Recycling

One way of trying to improve the results obtained from the Fogel-Huang algorithm is to recycle the data, using, in the case considered here, the Fogel-Huang ellipsoid after the 12th step as the initial ellipsoid for a new cycle and the same hyperplanes again, possibly as shifted according to the Belforte-Bona-Cerone method. Although this departs from the idea of updating the estimate of the feasible parameter set with each new data item, i.e., it is not an online method[3], it is applied to the data sets here. If, at a particular step, a pair of hyperplanes is found to be redundant, in that they do not cut the ellipsoid (more precisely, they either do not intersect the ellipsoid or they are tangent to it), that pair will be discarded[4].

However, although impressive compared with the Fogel-Huang algorithm without recycling, the results of doing this are not overly impressive in comparison with the minimum-volume ellipsoid characteristic lengths. As can be seen from Figures 3.44 to 3.45, most of the improvement is achieved in the first few cycles when the noise has a fairly peaked distribution, and in the first five cycles for the uniform distribution. After this cycle number, the characteristic length ratios settle down and are fairly distant from 1, which would correspond to the Fogel-Huang ellipsoid equalling the minimum-volume ellipsoid. For the uniform distribution, the mean improvement is shown in Table 3.3.

However, much of the improvement is for the worst cases for the single cycle algorithm, thus reducing the skewness of the results. In particular, the worst case for the singe cycle improves dramatically with recycling, with a final characteristic length which is close to the best case after 10 cycles.

Figures 3.23 to 3.25 illustrate that the improvement with dimension of the characteristic length persists when the data is recycled, but no real trend with regard to the change in the parameter-centre distance can be detected in Figures 3.26 to 3.28.

It turns out that a better idea is to utilise equation (3.16). Given that in recycling, in each cycle

[3]However, "onlineness" is a stretchable concept. "Windowing" methods, considering the N most recent observations could be online, so if the window here was 12 data items long, this would be an online method

[4]Although such redundant hyperplane pairs contain no "information" that is not contained in the other hyperplane pairs (they cannot intersect the feasible set except, possibly, in a degenerate vertex — one belonging to more than p hyperplanes), it is possible that retaining them might speed up the algorithm in exceptional cases, as they can be moved to be tangent to the ellipsoid *á la* Belforte-Bona-Cerone, and then, at a later stage, after approximations have been made to obtain later ellipsoids, they might cut the current ellipsoid, and it has not yet been shown that this cut cannot be "deep" enough to reduce the volume of that ellipsoid by using the Fogel-Huang algorithm. However, this is considered to be a somewhat unlikely event.

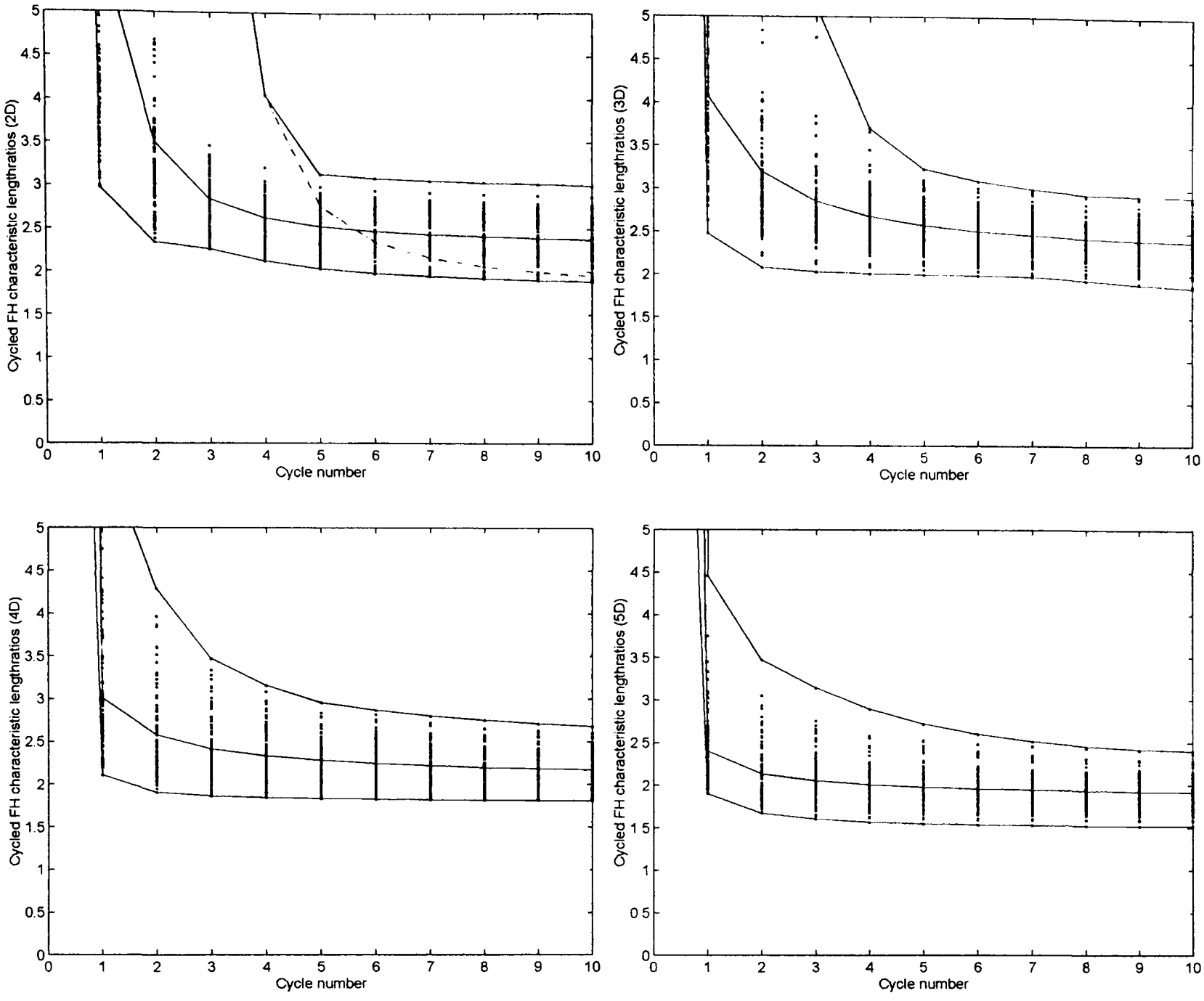


Figure 3.44: Characteristic length of Fogel-Huang ellipsoids (noise uniformly distributed). (The additional curve (dot-dash) gives the characteristic lengths of the "worst" data set of Figure 3.42.)

the entire set of available data is used, there is no further departure from "onlineness" involved in retaining all hyperplanes (again, possibly as shifted in previous steps) using equation (3.16) to choose the hyperplane pair producing the greatest reduction in the volume at each step. To make a comparison with recycling in fixed order, Table 3.3, the resulting characteristic length-ratio is examined after multiples of 12 steps in Table 3.4, in both cases for the data sets affected by uniformly distributed noise.

Comparing Figures 3.44 to 3.46 with Figures 3.47 to 3.49 reveals that, for each noise distribution, and for each dimension, choosing the hyperplane pairs at each step results in final minimum, mean and maximum final characteristic length-ratios which are better than the minimum, mean and maximum ratios for recycling in fixed order. These improvements are tabulated in Table 3.5 (Note that in the "maximum" and "minimum" rows in this Table, the comparison

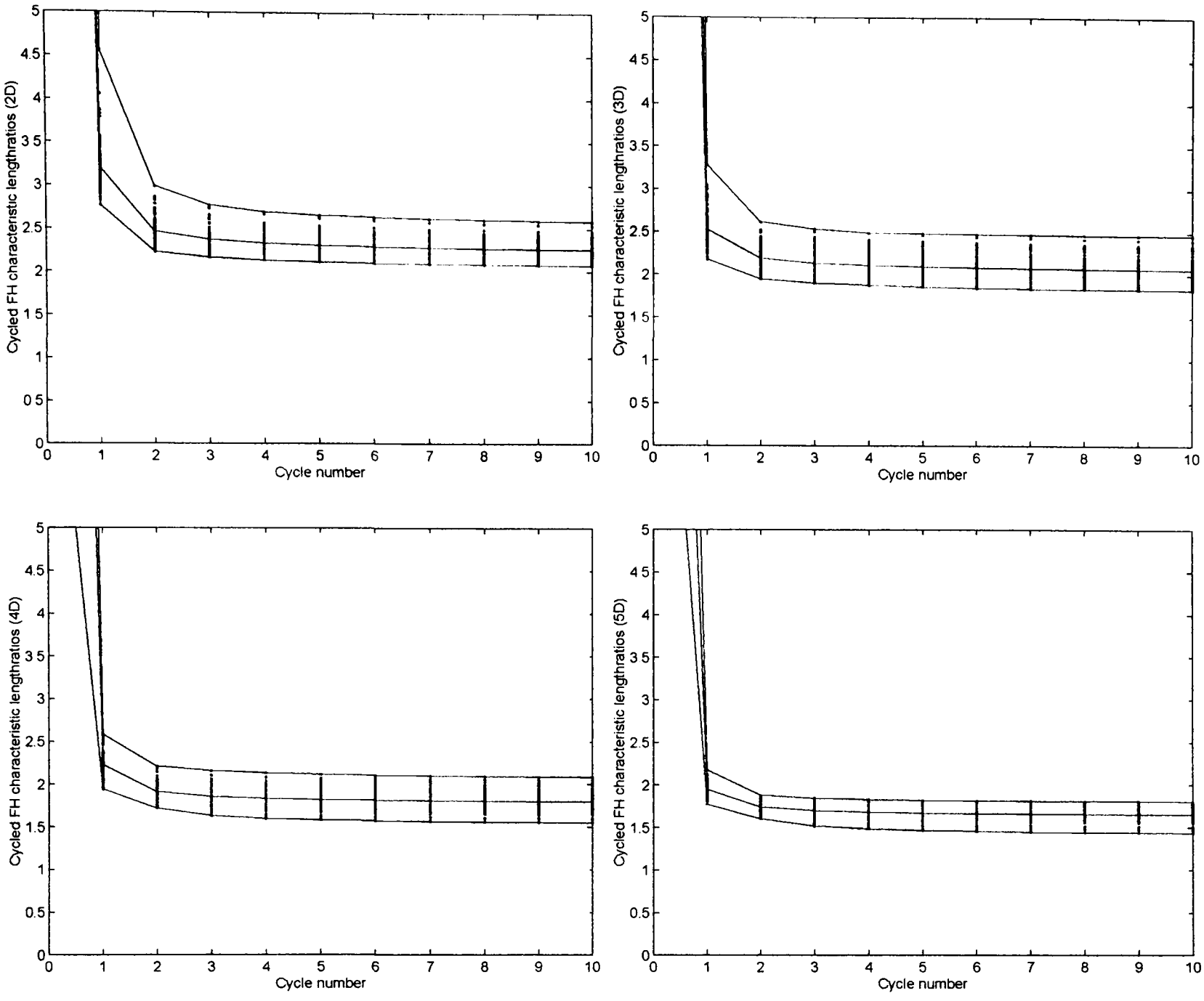


Figure 3.45: Characteristic length of Fogel-Huang ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

is not necessarily between the same data sets, as the worst performance, for example, may not be for the same data set for both algorithms). In addition, choosing the "best" pair results in a more rapid approach to the final value, which is almost achieved in most cases by the 24th step for the uniform error distribution and by the 12th step for the peaked distributions.

Figure 3.47 also reveals that the "worst" case for the single pass of the data using the Fogel-Huang algorithm rapidly approaches the "best" case for the "best pair" version — in fact, it is the 9th best out of the data sets for two dimensional parameter space derived using a uniform noise distribution.

In Table 3.6 the improvement in performance due to choosing the "best pair" over 12 steps, relative to the performance after the 12 steps of the single pass algorithm, is listed.

The behaviour of the "best pair" variant is interesting. It tends to settle on a three- or four-

Dimension	Mean characteristic length ratio			
	initial	after 1 cycle	after 2 cycles	after 10 cycles
2	166.9520	5.8258	3.4938	2.3751
3	106.8305	4.0756	3.2004	2.3903
4	64.3658	3.0071	2.5747	2.1707
5	40.6968	2.4012	2.1333	1.9185

Table 3.3: Improvement after recycling the data (uniform noise case).

Dimension	Mean characteristic length ratio			
	initial	after 12 steps	after 24 steps	after 120 steps
2	166.9520	2.5299	2.3421	2.2599
3	106.8305	2.8019	2.4756	2.2689
4	64.3658	2.5117	2.2965	2.1028
5	40.6968	2.2012	2.0402	1.8868

Table 3.4: Characteristic lengths after choosing the hyperplanes resulting in the greatest reduction in the characteristic length (uniform noise).

	Dimension							
	2		3		4		5	
σ_{noise}	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$
minimum	4.7%	24.3%	9.9%	1.3%	1.0%	1.3%	1.1%	1.2%
mean	4.8%	22.5%	5.1%	1.0%	3.1%	1.0%	1.6%	0.7%
maximum	8.8%	5.5%	3.9%	0.6%	6.2%	0.6%	2.4%	0.4%

Table 3.5: Improvement in characteristic length-ratios due to the choice of the "best" hyperplane pair over fixed order cycling (after 120 steps or 10 cycles — percentages relative to the length-ratio for fixed order recycling).

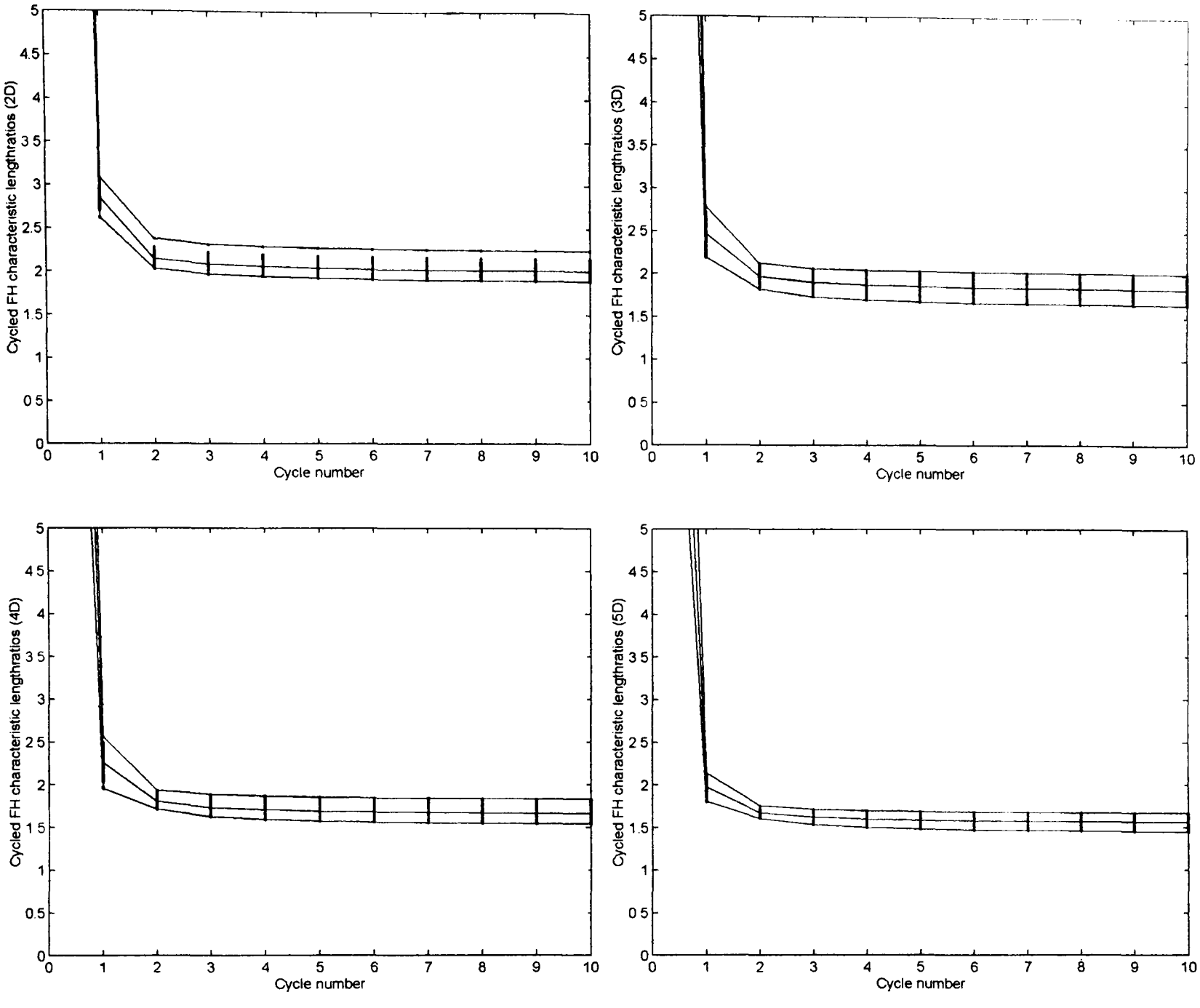
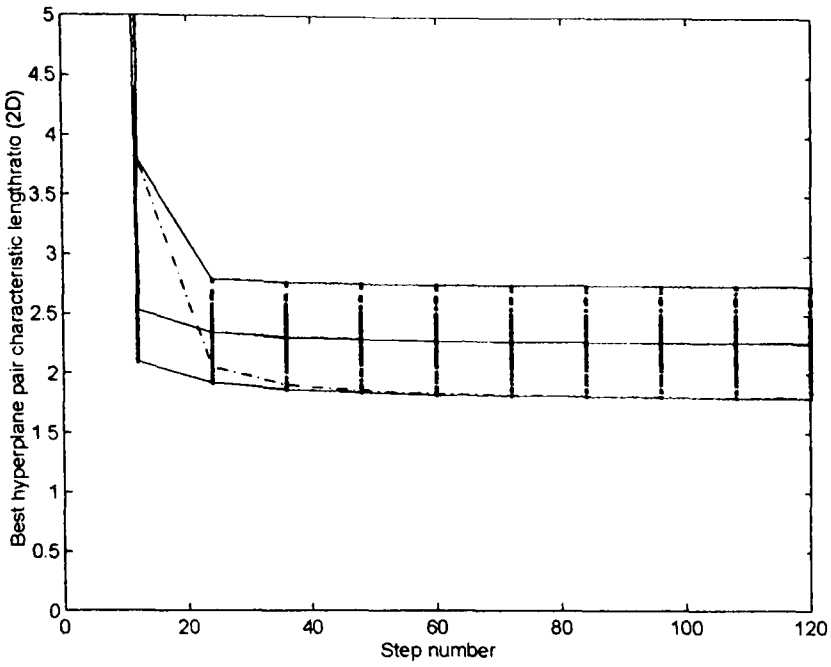
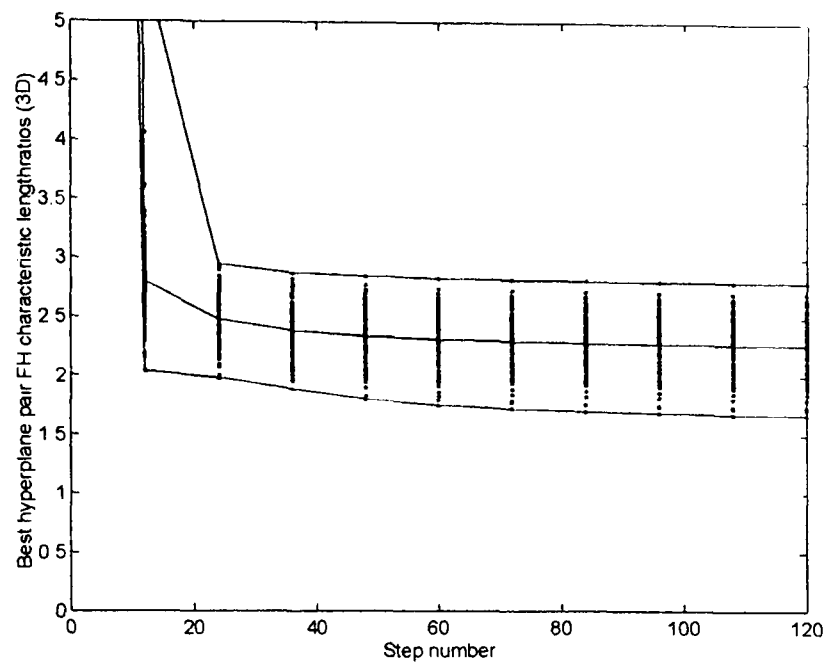


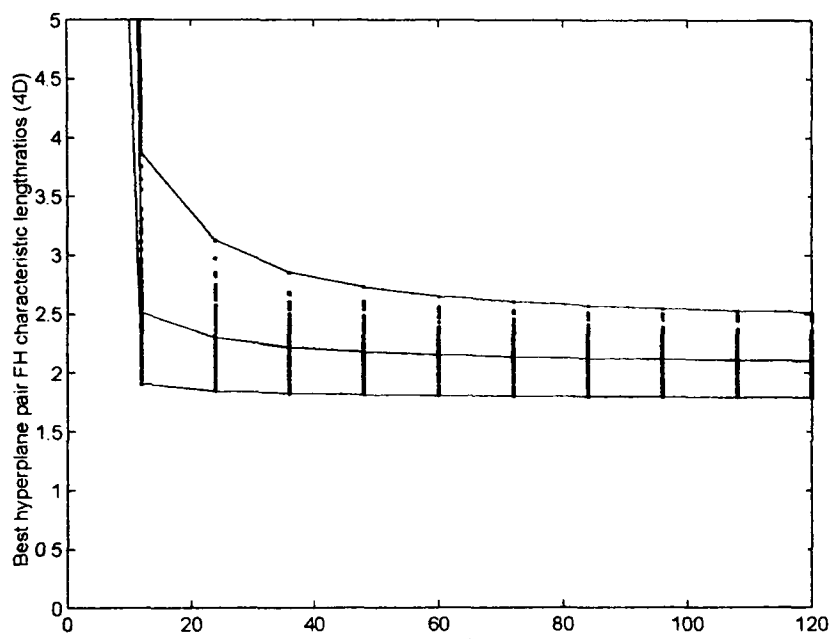
Figure 3.46: Characteristic length of Fogel-Huang ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

member subset of the available hyperplane pairs and cycle through that subset until it is "exhausted", when either a member is dropped, or a member added, or both, and then the new subset is cycled through.

However, it should not be thought that choosing the "best" hyperplane pair always results in a smaller ellipsoid than cycling in fixed order — it only "almost" always does. This is a manifestation of the following:







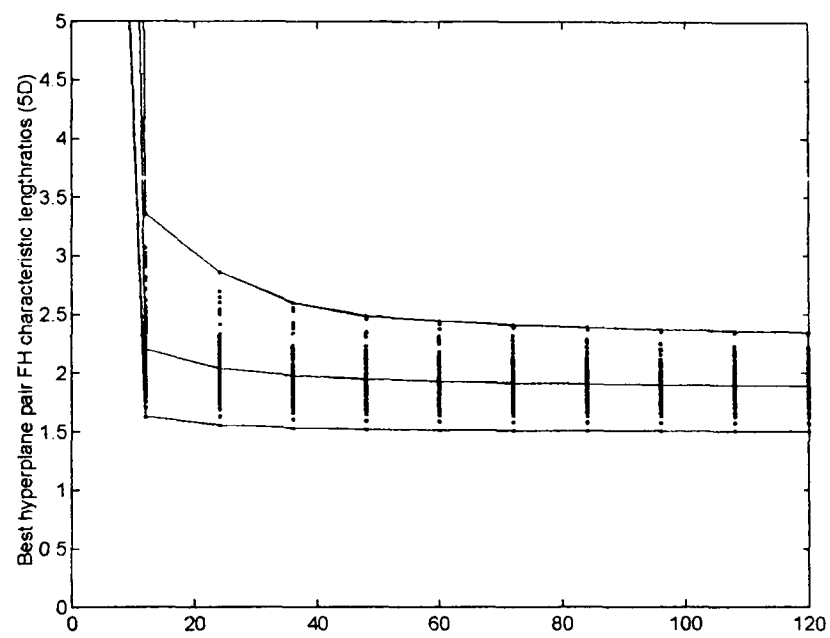


Figure 3.47: Characteristic length of Fogel-Huang ellipsoids, best hyperplane pair chosen at each step (uniformly distributed noise). (The additional curve again represents the "worst" data set for the unrecycled Fogel-Huang algorithm.)

Let $\mathcal{E}$ be an initial ellipsoid, and let $\Pi_1, \ldots, \Pi_t$ be strips between hyperplane pairs $\mathbb{H}_1^{\pm}, \ldots \mathbb{H}_t^{\pm}$. Define $\mathcal{E}_{k_1 \ldots k_s}$ to be the minimum-volume ellipsoid containing $\Pi_{k_s}^{s-1} \cap \mathcal{E}_{k_1 \ldots k_{s-1}}$, where $\Pi_{k_s}^{s-1}$ is the strip between the hyperplanes $\mathbb{H}_{k_s}^{\pm,s-1}$, which are the hyperplanes $\mathbb{H}_{k_s}^{\pm,s-2}$ possibly as shifted to be tangent to $\mathcal{E}_{k_1 \ldots k_{s-1}}$ according to the Belforte-Bona-Cerone method. Then, if $\mathrm{vol}(\mathcal{E}_{k_1 \ldots k_{s+1}}) = \min_{k \in \mathbb{N}_t}\{\mathrm{vol}(\mathcal{E}_{k_1 \ldots k_s k})\}$ and $\mathrm{vol}(\mathcal{E}_{k_1 \ldots k_{s+1} k_{s+2}}) = \min_{k \in \mathbb{N}_t}\{\mathrm{vol}(\mathcal{E}_{k_1 \ldots k_s k_{s+1} k})\}$, it is not necessarily the case that $\mathrm{vol}(\mathcal{E}_{k_1 \ldots k_{s+1} k_{s+2}}) < \mathrm{vol}(\mathcal{E}_{k_1 \ldots k_s k'_{s+1} k'_{s+2}(k'_{s+1})})$, where $\mathrm{vol}(\mathcal{E}_{k_1 \ldots k_s k'_{s+1} k'_{s+2}(k'_{s+1})}) = \min_{k \in \mathbb{N}_t}\{\mathrm{vol}(\mathcal{E}_{k_1 \ldots k_s k'_{s+1} k})\}$.

	Dimension							
	2		3		4		5	
σ_{noise}	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$
minimum	29.6%	41.8%	17.8%	22.5%	9.6%	17.4%	14.5%	14.5%
mean	56.6%	41.4%	31.2%	23.4%	16.5%	22.6%	8.3%	15.2%
maximum	96.3%	29.9%	58.2%	25.3%	35.0%	24.6%	24.8%	16.2%

Table 3.6: Improvement in characteristic length-ratios due to the choice of the "best" hyperplane pair over a single pass through the data using the Fogel-Huang algorithm (after 12 steps).

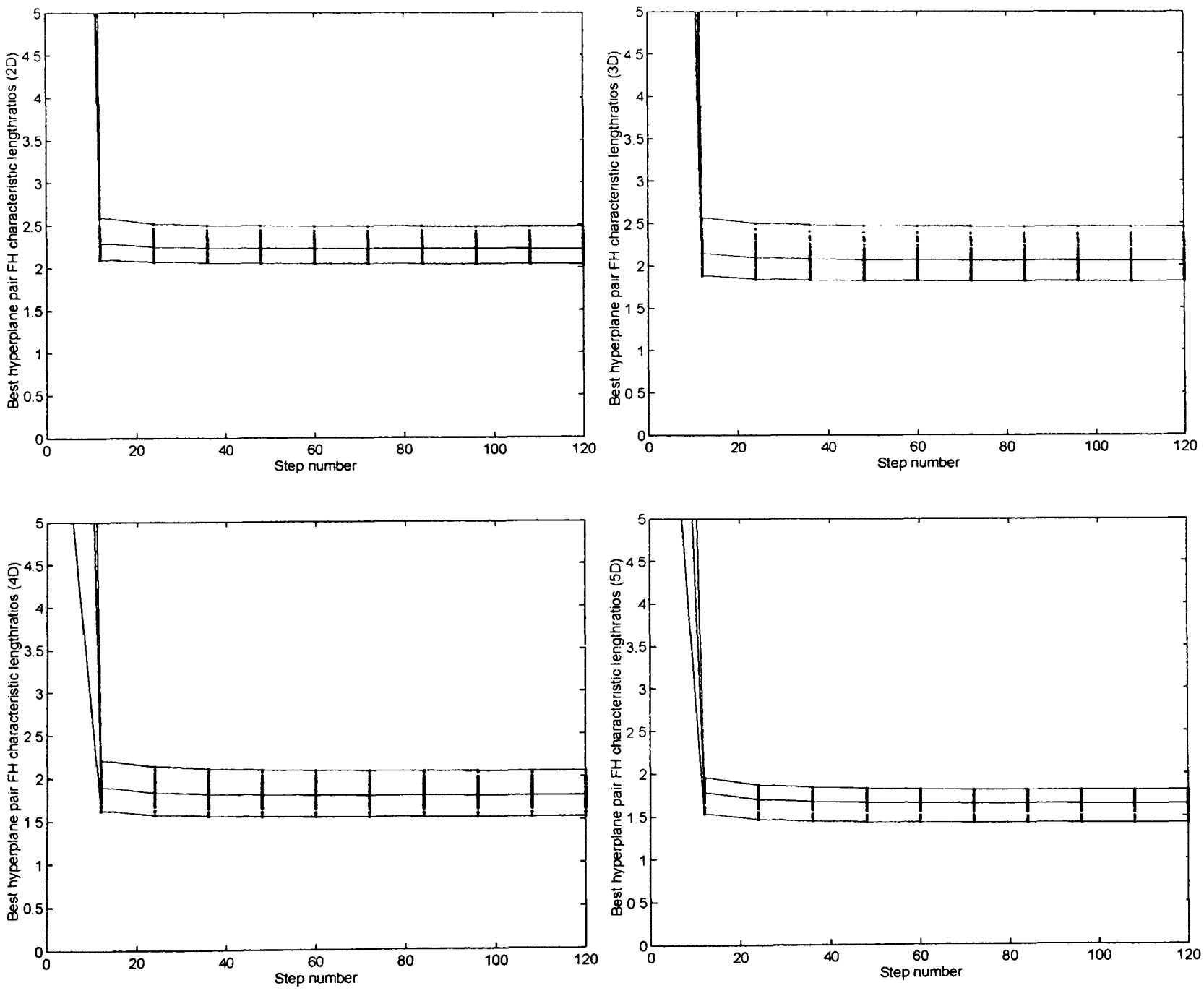


Figure 3.48: Characteristic length of Fogel-Huang ellipsoids, best hyperplane pair chosen at each step (normally distributed noise, $\sigma_t = 1/2\sqrt{3}$).

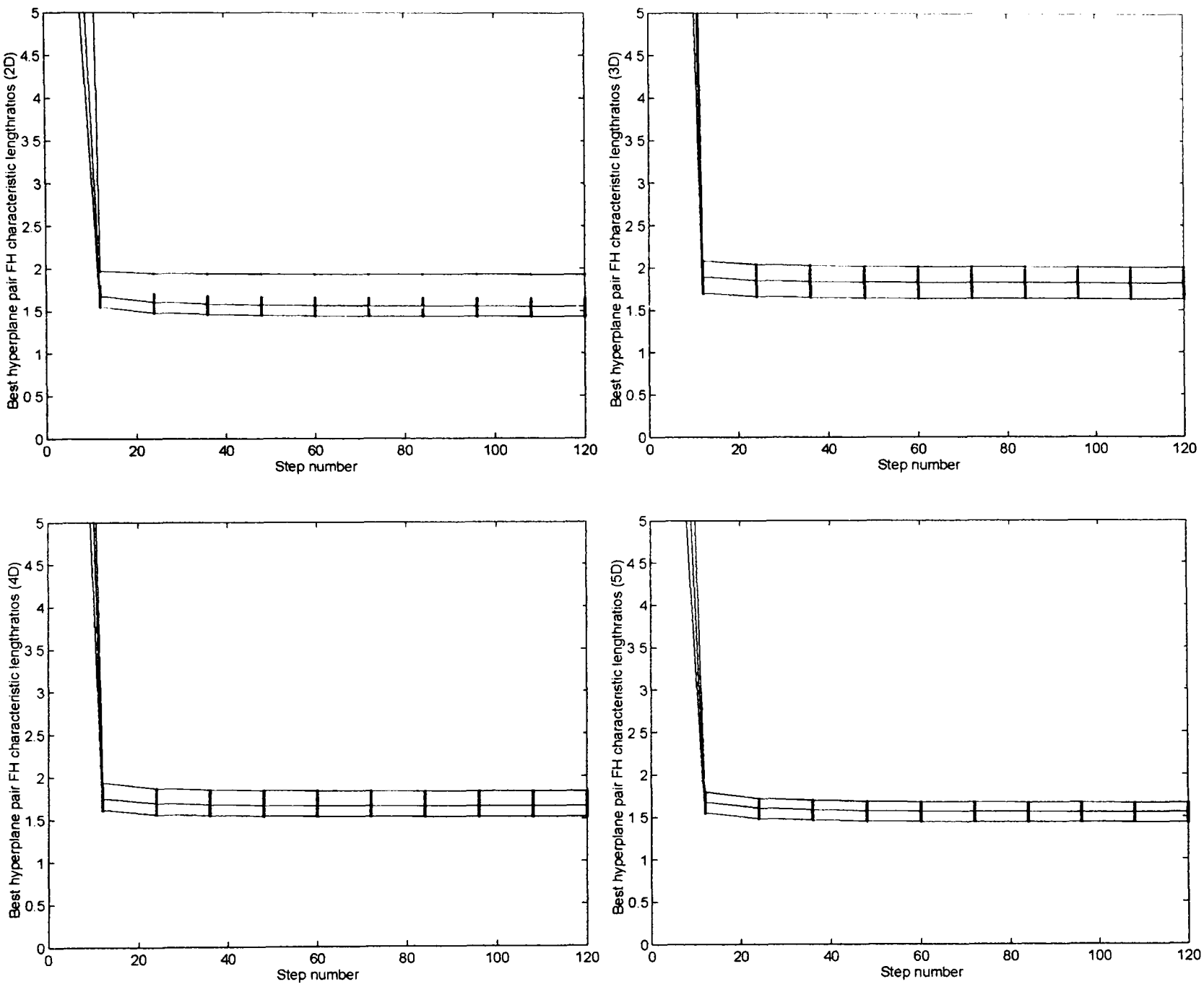


Figure 3.49: Characteristic length of Fogel-Huang ellipsoids, best hyperplane pair chosen at each step (normally distributed noise, $\sigma_t = 1/4\sqrt{3}$).

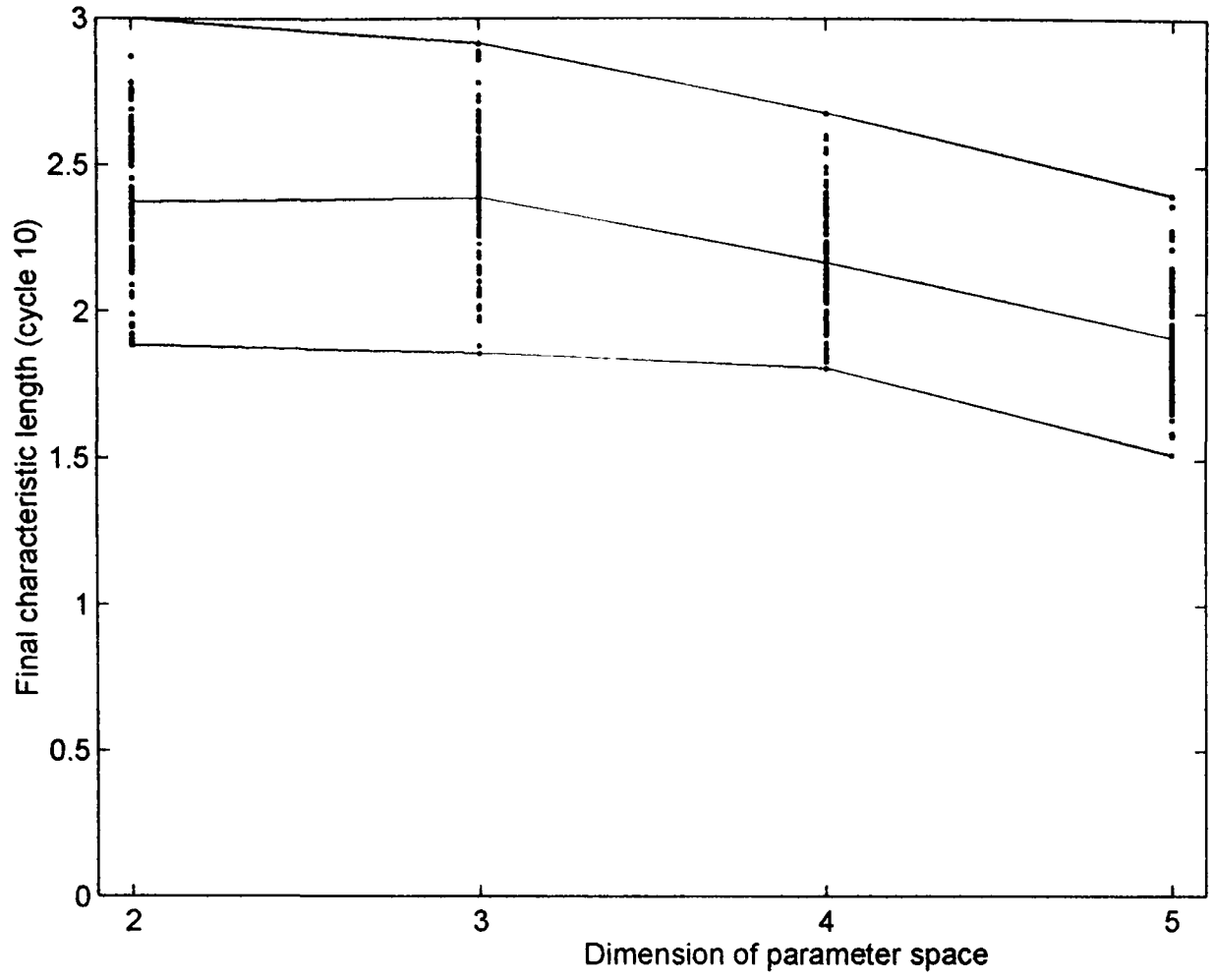


Figure 3.50: Final characteristic lengths for the Fogel-Huang algorithm (uniformly distributed noise).

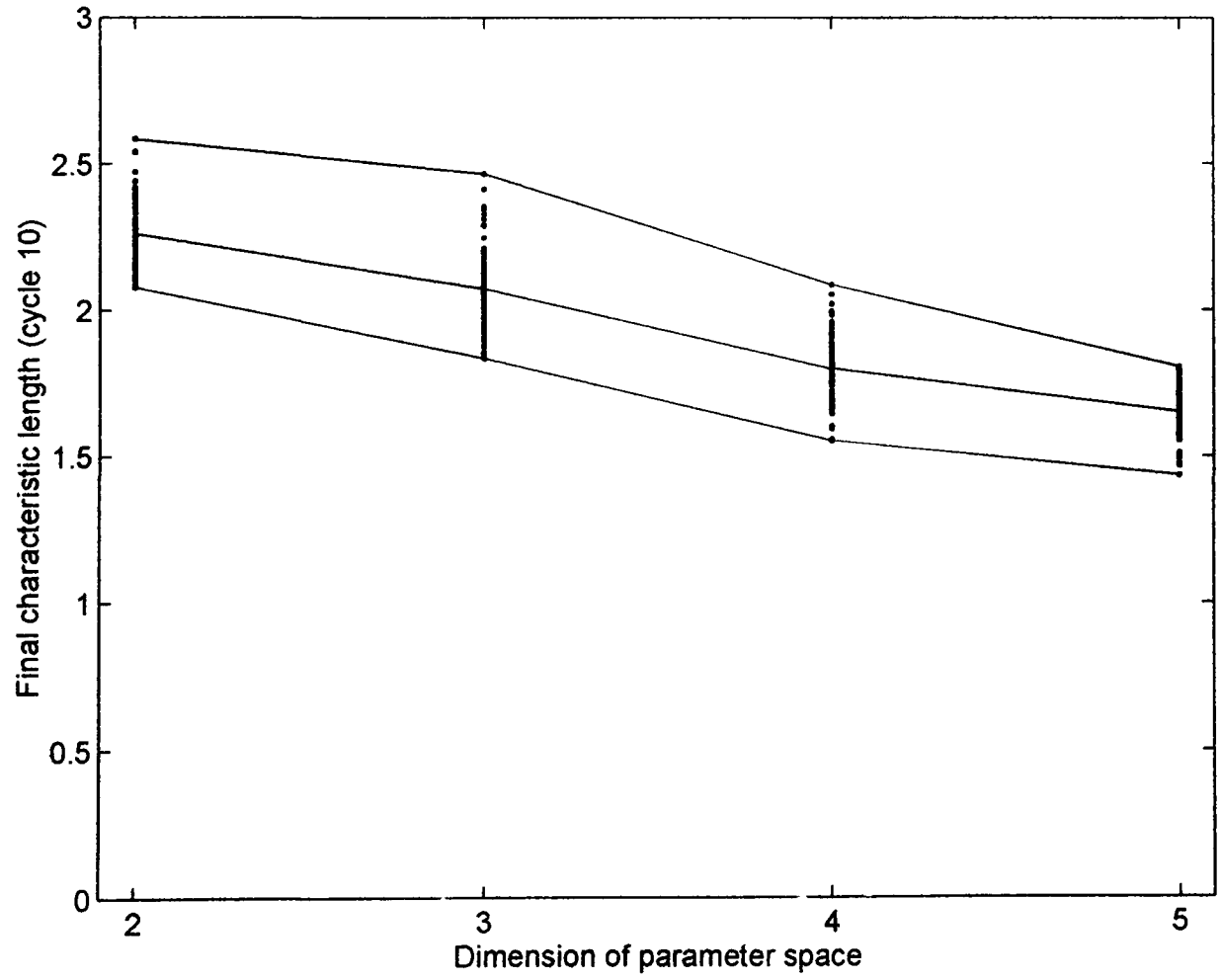


Figure 3.51: Final characteristic lengths for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

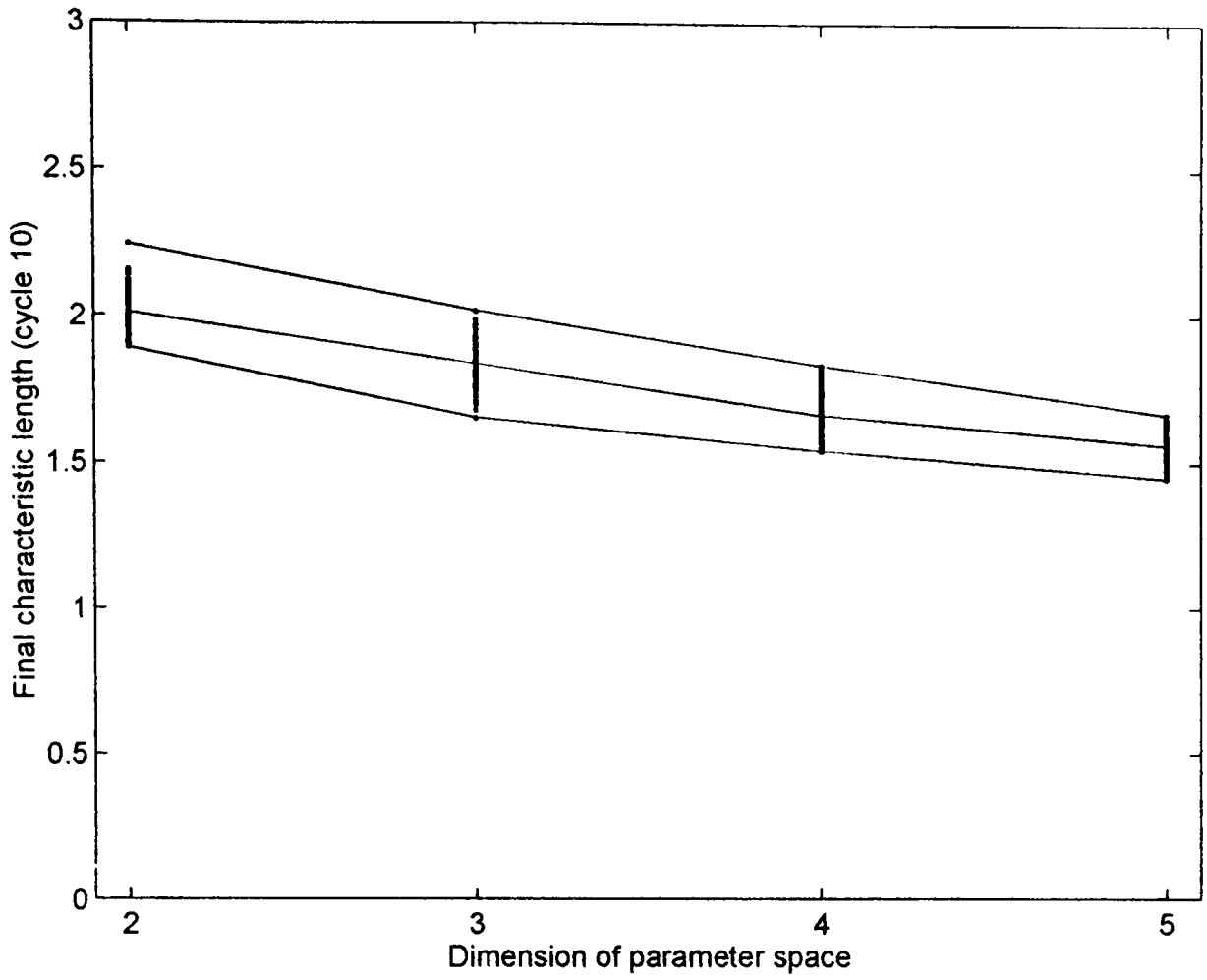


Figure 3.52: Final characteristic lengths for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

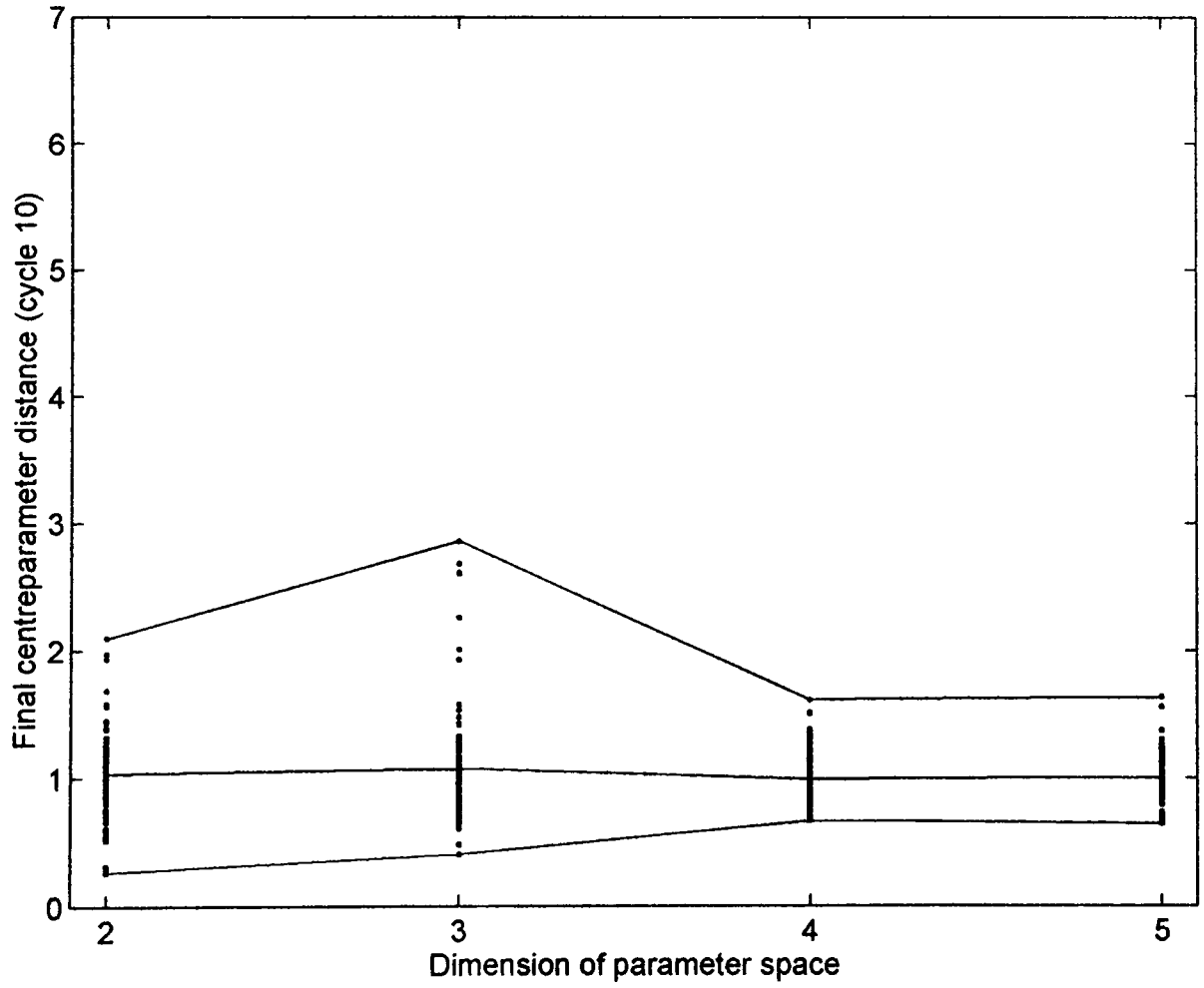


Figure 3.53: Final centre-parameter distance for the Fogel-Huang algorithm (uniformly distributed noise).

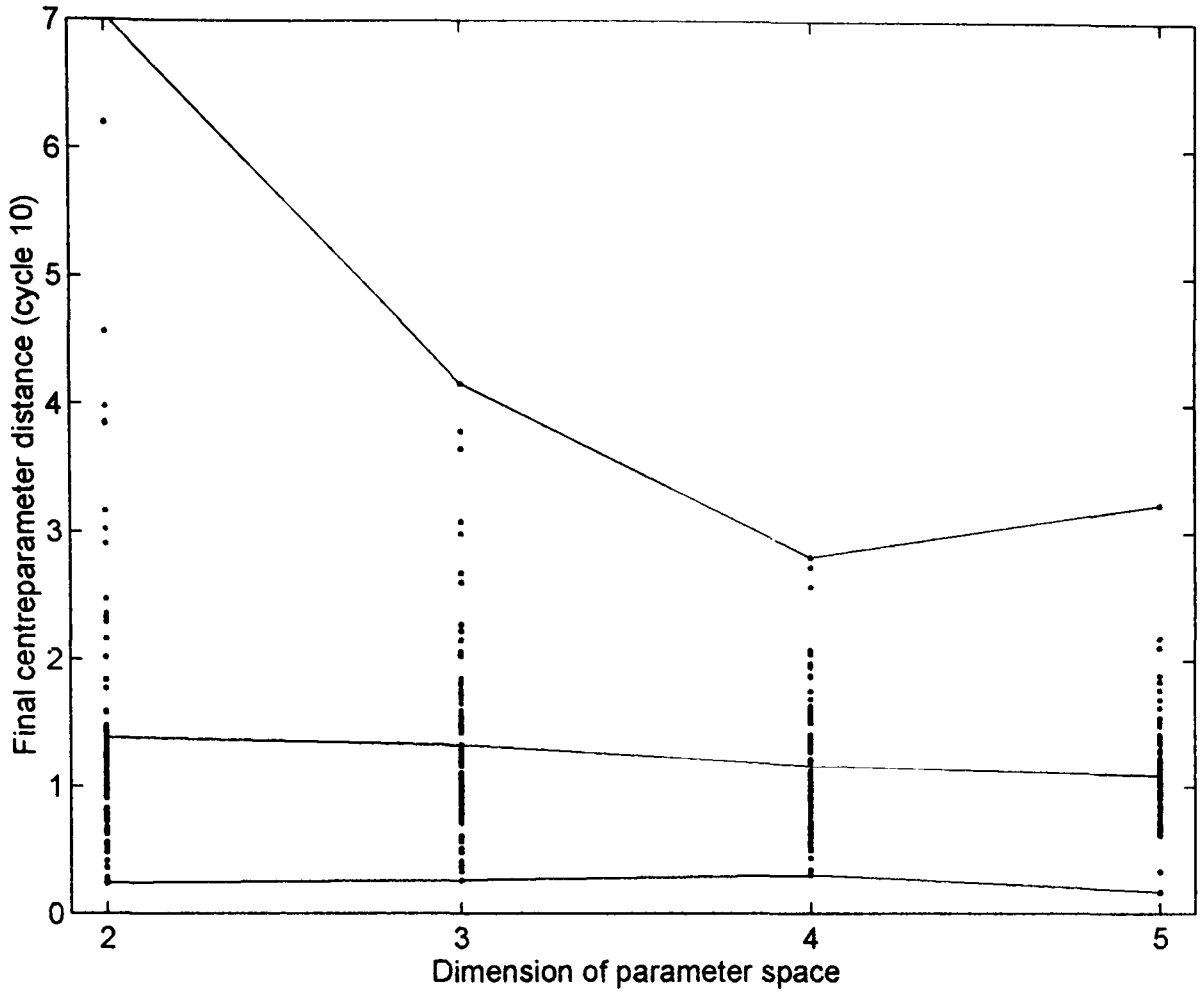


Figure 3.54: Final centre-parameter distance for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/2\sqrt{3}$).

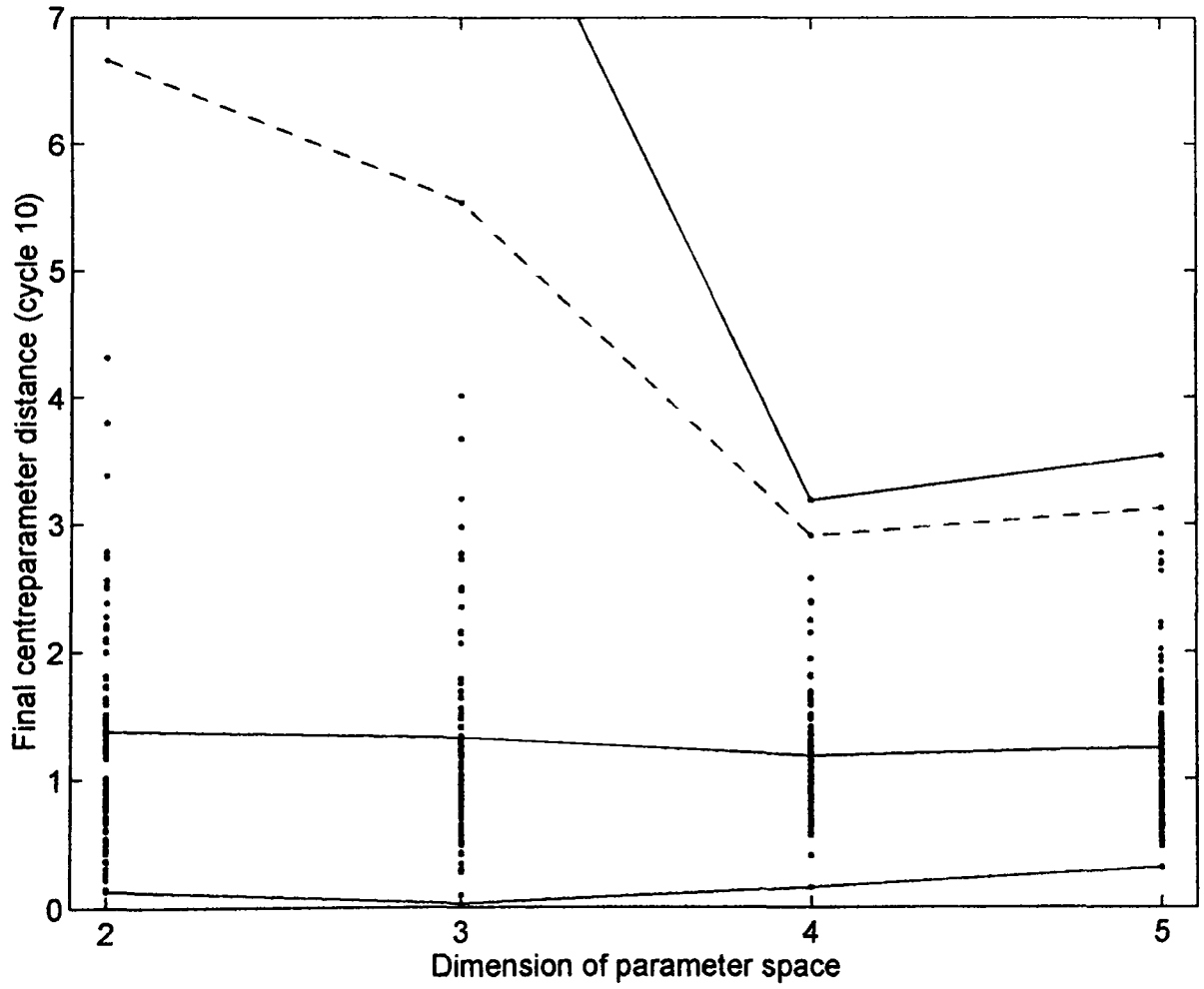


Figure 3.55: Final centre-parameter distance for the Fogel-Huang algorithm (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

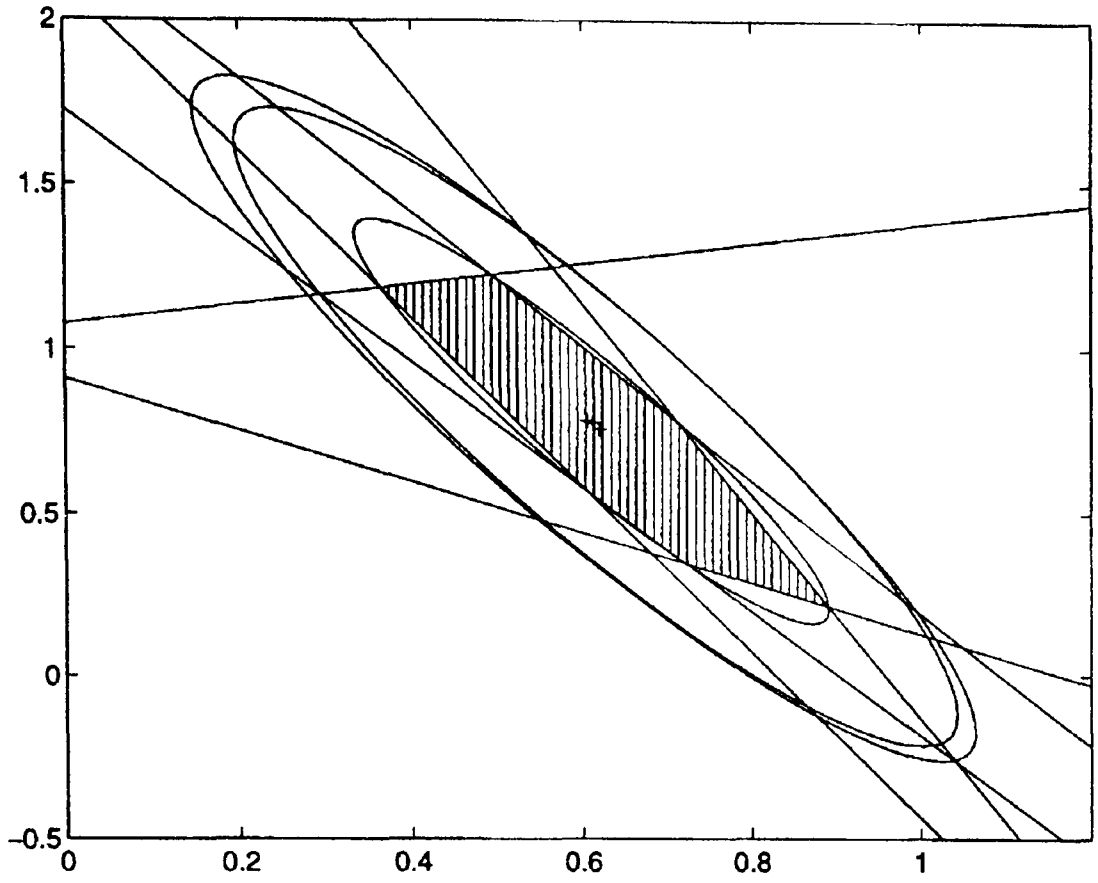


Figure 3.56: Fogel-Huang ellipsoids (for cycles 5 and 10), polytope (shaded) and BLJ ellipsoid for the two-dimensional data set leading to the greatest ratio between the volumes of the final Fogel-Huang ellipsoid and the BLJ ellipsoid.

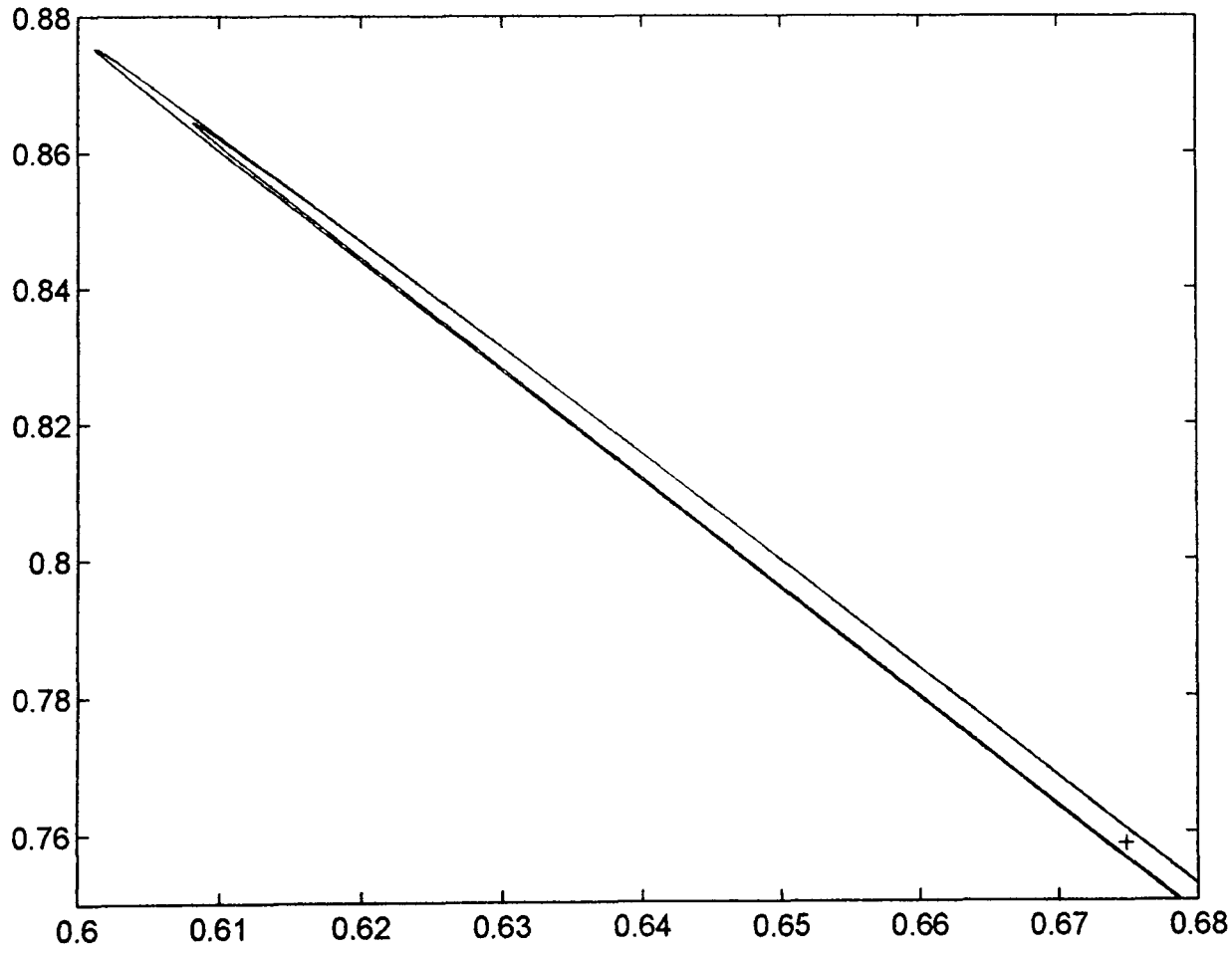


Figure 3.57: (Half of) Fogel-Huang after 10 cycles and BLJ ellipsoids for the data set of Figure 3.42 (the other halves look much the same). The ratio of the volumes is 1.2193.

3.3.4 Conclusions

In Table 3.7, a comparison is made between the empirical number of floating point operations (*flops*) needed to calculate the minimum-volume ellipsoid, the number of flops needed for 12 steps of the Fogel-Huang algorithm and the number of flops needed for 12 steps of the Fogel-Huang algorithm when the hyperplane pair leading to the greatest reduction in volume is selected at each step.

	Dimension							
	2		3		4		5	
Algorithm	mean	sd	mean	sd	mean	sd	mean	sd
Fogel-Huang	3.1	1.2	9.4	2.2	3.7	1.1	11	2.3
bFH	150	11	239	8.0	350	7.3	480	6.5
Minimum-volume	950	1200	5300	3500	750	560	34000	24000

Table 3.7: Mean and standard deviation of the number of kiloflops needed to calculate the minimum-volume ellipsoid, 12 steps of the Fogel-Huang algorithm and 12 steps of the Fogel-Huang algorithm selecting the "best" hyperplane pair at each step (bFH).

Although the number of flops required for the Fogel-Huang and minimum-volume ellipsoid algorithms for 4-dimensional parameter space appears to be anomalously low, the main feature to emerge from Table 3.7 is that the calculation of the minimum-volume ellipsoid is much more expensive than either of the other two algorithms. In addition, the standard deviations of the number of flops necessary for the minimum-volume algorithm is also much larger than the corresponding numbers for the other algorithms, even relative to the respective means. This is because number of operations demanded by the procedure for calculating the minimum-volume ellipsoid from the vertices of the contained polytope is very sensitive to the number of these vertices, whereas the time needed by the other two algorithms depends on whether or not the volume would be reduced at each step, and, in the case of the "best pair" variant, whether any redundant hyperplanes have been detected.

At any rate, the calculation of the minimum-volume ellipsoid is both much more expensive and less predictable in terms of cost.

The empirical average behaviour of the Fogel-Huang algorithm after 12 steps results in characteristic lengths roughly 2 to 5 times the minimum possible for an ellipsoid containing the final posterior set, and this means that it results in uncertainties in the parameters of about 2 to 5

times greater than does the minimum-volume ellipsoid.

However, the Fogel-Huang algorithm provides an online method of processing data as it arrives in a relatively rapid fashion, and is valuable for this reason.

Further, recycling the data, preferably not in fixed order, but by choosing the hyperplane pairs (either from all the data which has been encountered, or from the current "window") which result in the greatest reduction in volume at each step, is a fairly cheap way of improving the accuracy of the Fogel-Huang algorithm.

Nevertheless, the accuracy is still not all that great, and it is desirable to improve upon it without incurring too great a penalty in terms of additional computation.

Chapter 4

Modifications of the Fogel-Huang Algorithm

4.1 The Basic Modification

The idea behind the modification considered here is to use more than one hyperplane pair. The analogue of the Fogel-Huang procedure will then produce an ellipsoid containing the intersection of the regions contained between two or more hyperplane pairs and a preceding ellipsoid. The new ellipsoid, unlike that obtained for a single hyperplane pair, will not be the minimum volume ellipsoid containing the intersection. However, as the "information" contained in the hyperplane pairs other than the first is not entirely neglected during the operation, it seems likely that the resulting ellipsoid will be closer to the minimum-volume BLJ ellipsoid containing the intersection than is the ellipsoid obtained by utilising each hyperplane pair separately in a sequential application of the Fogel-Huang algorithm.

The intersection of the sets $\mathcal{E}$, Π_1 and Π_2, is considered, where $\mathcal{E}$ is defined by equation (3.1) and the Π_i by

$$x \in \Pi_i = \Pi(n_i, y_i) \Leftrightarrow (n_i^{\mathrm{T}} x - y_i)^2 \leq 1, i = 1, 2 \tag{4.1}$$

(as in equation (3.2)), with hyperplanes

$$x \in \mathbb{H}_{\pm}(n_i, y_i) \Leftrightarrow y_i - n_i^{\mathrm{T}} x = \pm 1. \tag{4.2}$$

The sum of $q_1 (\geq 0)$ times the inequality defining Π_1, $q_2 (\geq 0)$ times the inequality defining Π_2 and the inequality defining $\mathcal{E}$ is now found:

$$(x-a)^{\mathrm{T}} Q^{-1} (x-a) + q_1 (n_1^{\mathrm{T}} x - y_1)^2 + q_2 (n_2^{\mathrm{T}} x - y_2)^2 \leq 1 + q_1 + q_2. \tag{4.3}$$

Then

$$\begin{aligned} x^{\mathrm{T}}(Q^{-1}+q_1n_1n_1^{\mathrm{T}}+q_2n_2n_2^{\mathrm{T}})x-2x^{\mathrm{T}}(Q^{-1}a+y_1q_1n_1+y_2q_2n_2) & \\ +a^{\mathrm{T}}Q^{-1}a+y_1^2q_1+y_2^2q_2 &= \\ (x-\bar{a})^{\mathrm{T}}\tilde{Q}^{-1}(x-\bar{a})+a^{\mathrm{T}}Q^{-1}a+y_1^2q_1+y_2^2q_2-\bar{a}^{\mathrm{T}}\tilde{Q}^{-1}\bar{a} &\leq 1+q_1+q_2, \end{aligned} \tag{4.4}$$

where

$$\begin{aligned} \tilde{Q}=\tilde{Q}(q_1,q_2) &= (Q^{-1}+q_1n_1n_1^{\mathrm{T}}+q_2n_2n_2^{\mathrm{T}})^{-1} \\ &= \left[I_p-Q\frac{q_1(1+g_2q_2)n_1n_1^{\mathrm{T}}-hq_1q_2(n_1n_2{}^{\mathrm{T}}+n_2n_1{}^{\mathrm{T}})+q_2(1+g_1q_1)n_2n_2{}^{\mathrm{T}}}{1+g_1q_1+g_2q_2+(g_1g_2-h^2)q_1q_2}\right]Q, \end{aligned} \tag{4.5}$$

by two applications of the matrix inveision lemma, where $g_i=n_i^{\mathrm{T}}Qn_i$, $i=1,2$, $h=n_1^{\mathrm{T}}Qn_2$ and I_p is the p-dimensional identity matrix. Also,

$$\bar{a}=\bar{a}(q_1,q_2)=a+Q\frac{[\nu_1q_1(1+g_2q_2)-\nu_2hq_1q_2]n_1+[\nu_2q_2(1+g_1q_1)-\nu_1hq_1q_2]n_2}{1+g_1q_1+g_2q_2+(g_1g_2-h^2)q_1q_2}, \tag{4.6}$$

where $\nu_i=y_i-n_i^{\mathrm{T}}a$, $i=1,2$, so

$$-a^{\mathrm{T}}Q^{-1}a-y_1^2q_1-y_2^2q_2+\bar{a}^{\mathrm{T}}\tilde{Q}^{-1}\bar{a}=-\frac{\nu_1^2q_1(1+g_2q_2)-2\nu_1\nu_2hq_1q_2+\nu_2^2q_2(1+g_1q_1)}{1+g_1q_1+g_2q_2+(g_1g_2-h^2)q_1q_2}.$$

Hence

$$(x-\bar{a}(q_1,q_2))^{\mathrm{T}}\bar{Q}(q_1,q_2)^{-1}(x-\bar{a}(q_1,q_2))\leq 1 \tag{4.7}$$

where

$$\bar{Q}(q_1,q_2)=\left[1+q_1+q_2-\frac{\nu_1^2q_1(1+g_2q_2)-2\nu_1\nu_2hq_1q_2+\nu_2^2q_2(1+g_1q_1)}{1+g_1q_1+g_2q_2+(g_1g_2-h^2)q_1q_2}\right]\tilde{Q}(q_1,q_2), \tag{4.8}$$

defines a family of ellipsoids $\mathcal{E}(\bar{a}(q_1,q_2),\bar{Q}(q_1,q_2))$ each member of which is guaranteed to contain $\mathcal{E}(a,Q)\cap\Pi_1\cap\Pi_2$.

The volume of $\mathcal{E}(\bar{a}(q_1,q_2),\bar{Q}(q_1,q_2))$ is proportional to

$$\begin{aligned} \det\bar{Q}(q_1,q_2) &= \left[1+q_1+q_2-\frac{\nu_1^2q_1(1+g_2q_2)-2\nu_1\nu_2hq_1q_2+\nu_2^2q_2(1+g_1q_1)}{1+g_1q_1+g_2q_2+(g_1g_2-h^2)q_1q_2}\right]^p\det\tilde{Q}(q_1,q_2) \\ &= \frac{n(q_1,q_2)^p}{d(q_1,q_2)^{p+1}}\det Q \end{aligned} \tag{4.9}$$

after using

$$\det(I_p+uv^{\mathrm{T}})=1+v^{\mathrm{T}}u$$

twice, where

$$\begin{aligned} d(q_1,q_2) &= 1+g_1q_1+g_2q_2+(g_1g_2-h^2)q_1q_2 \qquad (4.10)\\ n(q_1,q_2) &= (1+q_1+q_2)d(q_1,q_2)\\ &\qquad -\nu_1^2q_1(1+g_2q_2)+2\nu_1\nu_2hq_1q_2-\nu_2^2q_2(1+g_1q_1)\\ &= 1+(1+g_1-\nu_1^2)q_1+(1+g_2-\nu_2^2)q_2\\ &\quad +g_1q_1^2+(g_1+g_2+g_1g_2-h^2-\nu_1^2g_2+2\nu_1\nu_2h-\nu_2^2g_1)q_1q_2+g_2q_2^2\\ &\quad +(g_1g_2-h^2)(q_1+q_2)q_1q_2. \qquad (4.11) \end{aligned}$$

To minimise the volume of $\mathcal{E}(\bar{a}(q_1,q_2),\bar{Q}(q_1,q_2))$, we need to minimise $n(q_1,q_2)^p/d(q_1,q_2)^{p+1}$. If we set the derivatives with respect to q_1 and q_2 of this last expression to zero, we need to find q_1 and q_2 such that

$$pd(q_1,q_2)\frac{\partial n(q_1,q_2)}{\partial q_i}-(p+1)n(q_1,q_2)\frac{\partial d(q_1,q_2)}{\partial q_i}=0, \quad i=1,2 \qquad (4.12)$$

as $n(q_1,q_2)$ cannot be zero for non-degenerate ellipsoids. One of equations (4.12) is quadratic in q_1 and cubic in q_2, the other is cubic in q_2 and quadratic in q_2. If we use the equation which is quadratic in q_1 to find an expression for q_1 and substitute the result in the other equation, and then separate the terms which involve the (single) square root, we will be able to square both sides and obtain a polynomial in q_2. Using Maple© (see Appendix) to do this, we find that the resulting polynomial is of degree 5. Each root of this polynomial leads, in general, to two different q_1, so we obtain 10 different pairs (q_1,q_2), each of which corresponds to a possible minimum of the volume (clearly, many of the pairs will be spurious). However, we can immediately eliminate any pair which does not consist of two nonnegative reals.

If $p=2$, there is a complicating factor: the minimum volume ellipse containing $\mathcal{E}(a,Q)\cap\Pi_1\cap\Pi_2$ may be the minimum volume ellipse containing $\Pi_1\cap\Pi_2$.

4.1.1 Behaviour of $\mathcal{E}(\bar{a}(q_1,q_2),\bar{Q}(q_1,q_2))$ in Two Dimensions as $q_1,q_2\to\infty$

We are interested in this behaviour for two reasons: as $q_1,q_2\to\infty$ it seems reasonable from inequality (4.3) that $\mathcal{E}(\bar{a}(q_1,q_2),\bar{Q}(q_1,q_2))$ would tend to an ellipse containing $\Pi_1\cap\Pi_2$ which is independent of Q and we wish to investigate whether this is so; we are also concerned about numerical problems in the basic equation (4.8) of the modified algorithms when the second

term in square brackets in equation (4.5) is close in value to the identity matrix, which, as we will see, can occur when both q_1 and q_2 are large.

Let

$$R := Q\left(g_2 n_1 {n_1}^{\mathrm{T}} - h(n_1 {n_2}^{\mathrm{T}} + n_2 {n_1}^{\mathrm{T}}) + g_1 n_2 {n_2}^{\mathrm{T}}\right).$$

Then ${n_i}^{\mathrm{T}} R n_j = (g_1 g_2 - h^2){n_i}^{\mathrm{T}} n_j, i, j \in \{1, 2\}$, which, in two dimensions, is enough to prove that $R = (g_1 g_2 - h^2) I_2$ (provided that n_1 and n_2 are linearly independent, which we assume anyway.

Hence

$$\tilde{Q}(q_1, q_2) = \left[I_2 - \frac{(g_1 g_2 - h^2) I_2 + q_2^{-1} Q n_1 {n_1}^{\mathrm{T}} + q_1^{-1} Q n_2 {n_2}^{\mathrm{T}}}{g_1 g_2 - h^2 + g_2 q_1^{-1} + g_1 q_2^{-1} + q_1^{-1} q_2^{-1}}\right] Q$$

(where, of course, this holds only if $p = 2$), or

$$\begin{aligned} \tilde{Q}(q_1, q_2) &= \frac{1}{g_1 g_2 - h^2}\left[q_1^{-1}(g_2 I_2 - Q n_2 {n_2}^{\mathrm{T}}) + q_2^{-1}(g_1 I_2 - Q n_1 {n_1}^{\mathrm{T}})\right] Q \\ &\quad + \mathrm{O}(\max_{i=1,2} q_i^{-2}) \end{aligned} \tag{4.13}$$

on expanding about $(q_1, q_2) = (\infty, \infty)$.

By equation (4.8),

$$\begin{aligned} \bar{Q}(q_1, q_2) &= \frac{1}{g_1 g_2 - h^2}\left[\left((g_1 + g_2) I_2 - Q(n_1 {n_1}^{\mathrm{T}} + n_2 {n_2}^{\mathrm{T}})\right) Q\right. \\ &\quad \left. + \frac{q_1}{q_2}(g_1 I_2 - Q n_1 {n_1}^{\mathrm{T}}) Q + \frac{q_2}{q_1}(g_2 I_2 - Q n_2 {n_2}^{\mathrm{T}}) Q\right] \\ &\quad + \mathrm{O}(\max\{q_1 q_2^{-2}, q_1^{-2} q_2\}) \end{aligned} \tag{4.14}$$

and specialising to the case where $q_2 \to k q_1$ as $q_1 \to \infty$, where $k \in [0, \infty]$ rather than $q_1^{-1} q_2$ oscillating, for example),

$$\bar{Q}(q_1, q_2) \to \frac{[(1 + k^{-1}) g_1 + (1 + k) g_2] I_2 - Q[(1 + k^{-1}) n_1 {n_1}^{\mathrm{T}} + (1 + k) n_2 {n_2}^{\mathrm{T}}]}{g_1 g_2 - h^2} Q.$$

Using Maple©, we find that this expression is

$$\frac{1 + k}{k} \frac{(|n_1|^2 + k|n_2|^2) I_2 - n_1 {n_1}^{\mathrm{T}} - k n_2 {n_2}^{\mathrm{T}}}{|n_1 \times n_2|^2}, \tag{4.15}$$

where "$\times$" is the vector cross-product in three dimensions (the third dimension being present here purely for the definition of the cross-product!). This expression has the determinant

$$\frac{(1 + k)^2}{k} \frac{1}{|n_1 \times n_2|^2} \tag{4.16}$$

which has a minimum when $k = 1$, so $\mathcal{E}(\bar{a}(q_1, q_2), \bar{Q}(q_1, q_2))$ approaches an ellipse with matrix

$$\frac{2}{|n_1 \times n_2|^2}\left[(|n_1|^2 + |n_2|^2) I_2 - n_1 {n_1}^{\mathrm{T}} - n_2 {n_2}^{\mathrm{T}}\right] \tag{4.17}$$

when $q_1, q_2 \to \infty$ along $q_2 = q_1$ and this ellipse has the minimum volume among those approached as $q_1, q_2 \to \infty$ along $q_2 = kq_1$.

From equation (4.6)

$$\begin{aligned}
\bar{a}(q_1, q_2) &\to a + Q\frac{(g_2\nu_1 - h\nu_2)n_1 + (g_1\nu_2 - h\nu_1)n_2}{g_1g_2 - h^2} \\
&= Q\frac{(g_2y_1 - hy_2)n_1 + (g_1y_2 - hy_1)n_2}{g_1g_2 - h^2} \\
&\qquad + \left[I_2 - Q\frac{g_2n_1n_1{}^{\mathrm{T}} - h(n_1n_2{}^{\mathrm{T}} + n_2n_1{}^{\mathrm{T}}) + g_1n_2n_2{}^{\mathrm{T}}}{g_1g_2 - h^2}\right]a \\
&= Q\frac{(g_2y_1 - hy_2)n_1 + (g_1y_2 - hy_1)n_2}{g_1g_2 - h^2}
\end{aligned}$$

as $q_1, q_2 \to \infty$ along any path in the q_1q_2-plane. By Maple© calculations, this is

$$\bar{a}(q_1, q_2) \to \frac{n_1n_2{}^{\mathrm{T}} - n_2n_1{}^{\mathrm{T}}}{|n_1|^2|n_2|^2 - (n_1{}^{\mathrm{T}}n_2)^2}\,(y_1n_2 - y_2n_1) =: a_{(2)}. \tag{4.18}$$

It can be shown, by making a (nonsingular) linear coordinate transformation such that the n_i are parallel to the coordinate axes and are of length 1, that $a_{(2)}$ is the centre of every member of the family of ellipses each member of which has a boundary passing through the four points of $\Pi_1 \cap \Pi_2$, and that the matrices of the entire family are given by

$$\frac{2}{|n_1|^2|n_2|^2 - (n_1{}^{\mathrm{T}}n_2)^2}\,\frac{1}{(1-\kappa^2)}[\{(1+\kappa)|n_1|^2 + (1-\kappa)|n_2|^2\}I_2 - (1+\kappa)n_1n_1{}^{\mathrm{T}} - (1-\kappa)n_2n_2{}^{\mathrm{T}}],$$

for $\kappa \in (-1, 1)$. When κ is set equal to $(1-k)/(1+k)$, this reduces to the right-hand side of equation (4.15), so that the ellipses $\mathcal{E}(\bar{a}(q_1, q_2), \bar{Q}(q_1, q_2))$ obtained by letting $(q_1, q_2) \to (\infty, \infty)$ along $q_2 = kq_2$, $k \in (0, \infty$, are in 1-1 correspondence with the entire family of ellipses whose boundaries contain $\Pi_1 \cap \Pi_2$.

Thus we have shown that when $(q_1, q_2) \to (\infty, \infty)$ along $q_2 = kq_1$, $\mathcal{E}(\bar{a}(q_1, q_2), \bar{Q}(q_1, q_2))$ tends to a member, dependent on k but independent of Q and a, of the family of ellipses containing the boundary of $\Pi_1 \cap \Pi_2$, and because this family has as a member the minimum-area ellipse containing (for $\kappa = 0$), $\mathcal{E}(\bar{a}(q_1, q_2), \bar{Q}(q_1, q_2))$ tends to that minimum-area ellipse, the Behrend-Löwner ellipse (for $k = 0$). We have also shown that when $\det \tilde{Q}(q_1, q_2)Q^{-1}$ is "small" (as calculated according to equation (4.5)), so that we are likely to have numerical problems, we can replace the right-hand side of equation (4.8) by the right-hand side of equation (4.15), with $k = q_2/q_1$.

4.2 Two Pairs at a Time

4.2.1 Hyperplane Shifting — Two Pairs at a Time

Suppose we have two hyperplane pairs, $\mathbb{H}_1^{\pm}(n_1, y_1)$, $\mathbb{H}_2^{\pm}(n_2, y_2)$ and an ellipsoid $\mathcal{E} = \mathcal{E}(a, Q)$, such that $\mathcal{E} \cap \Pi_1(n_1, y_1) \cap \Pi_2(n_2, y_2) \neq \emptyset$. Our task here is to find the narrowest strip $\Pi_2'(n_2', y_2')$ bounded by hyperplanes $\mathbb{H}_2'^{\pm}$ parallel to $\mathbb{H}_2^{\pm}$ which contains $\mathcal{E} \cap \Pi_1 \cap \Pi_2$. As we already know that n_2' is proportional to n_2, this is equivalent to finding points $x^+ \in \mathbb{H}_2'^{+}$ and $x^- \in \mathbb{H}_2'^{-}$ which will then completely determine Π_2'.

Let us concentrate on $\mathbb{H}_2^+$. If $\mathbb{H}_2^+$ intersects $\mathcal{E} \cap \Pi_1$, $\mathbb{H}_2'^{+} = \mathbb{H}_2^+$ and we can put $x^+ = (y_2 + 1)n_2/{n_2}^{\mathrm{T}} n_2 \in \mathbb{H}_2^+$.

Suppose $\mathbb{H}_2^+ \cap \mathcal{E} \cap \Pi_1 = \emptyset$. If $n_2 \propto n_1$, we may simply choose the two hyperplanes from the set $\{\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-\}$ which bound the narrowest strip, so we assume that $\{n_1, n_2\}$ is a linearly independent set.

Initially, we suppose that $\mathbb{H}_2^+$ does not intersect $\mathcal{E}$. This will be the case if $\max_{x \in \mathcal{E}} {n_2}^{\mathrm{T}} x$ is less than the value of ${n_2}^{\mathrm{T}} x$ on $\mathbb{H}_2^+$, which is $y_2 + 1$. But ${n_2}^{\mathrm{T}} x$ achieves its maximum value on $\mathcal{E}$ when $x = x_{\mathcal{E}}^+ = a + \frac{1}{\sqrt{g_2}} Q n_2$ (this can be shown by utilising the Lagrangian (see Goldstein [10]) $L = {n_2}^{\mathrm{T}} x - \lambda((x-a)^{\mathrm{T}} Q^{-1}(x-a) - 1))$, and this value is ${n_2}^{\mathrm{T}} a + \sqrt{g_2}$. But $y_2 + 1 > {n_2}^{\mathrm{T}} a + \sqrt{g_2}$ is equivalent to $\nu_2 + 1 > \sqrt{g_2}$.

If $x_{\mathcal{E}}^+$ is contained in Π_1, then $({n_1}^{\mathrm{T}} x_{\mathcal{E}} - y_1)^2 \leq 1$ or $\left(\frac{h}{\sqrt{g_2}} - \nu_1\right)^2 \leq 1$, and we may shift $\mathbb{H}_2^+$ to pass through $x_{\mathcal{E}}^+$, by setting $x^+ = x_{\mathcal{E}}^+$.

This will be, of course, equivalent to making the transformations (3.13) and (3.14) or the transformation (3.15).

If, on the other hand, $\left(\frac{h}{\sqrt{g_2}} - \nu_1\right)^2 > 1$, then we may shift $\mathbb{H}_2^+$ so that it touches $\mathcal{E} \cap \Pi_1$, and we may also do this even if $\mathbb{H}_2^+$ intersects $\mathcal{E}$.

We are now interested in finding $x_{\Pi_1 \mathcal{E}}^+$ such that ${n_2}^{\mathrm{T}} x_{\Pi_1 \mathcal{E}}^+ = \max_{x \in \mathcal{E} \cap \Pi_1} {n_2}^{\mathrm{T}} x = \max\{{n_2}^{\mathrm{T}} x_{\mathbb{H}_1^+ \mathcal{E}}^+, {n_2}^{\mathrm{T}} x_{\mathbb{H}_1^- \mathcal{E}}^+\}$, where $x_{\mathbb{H}_1^+ \mathcal{E}}^+$ (respectively $x_{\mathbb{H}_1^- \mathcal{E}}^+$) is such that ${n_2}^{\mathrm{T}} x_{\mathbb{H}_1^+ \mathcal{E}}^+ = \max_{x \in \mathcal{E} \cap \mathbb{H}_1^+} {n_2}^{\mathrm{T}} x$ (respectively ${n_2}^{\mathrm{T}} x_{\mathbb{H}_1^- \mathcal{E}}^+ = \max_{x \in \mathcal{E} \cap \mathbb{H}_1^-} {n_2}^{\mathrm{T}} x$). To find $x_{\mathbb{H}_1^+ \mathcal{E}}^+$ we will need the following Theorem:

Theorem 4.1: Let $n \in \mathbb{R}^p - \{0\}$, and let $\{n_1, \ldots, n_s\} \in (\mathbb{R}^p)^s$ be such that $\{n, n_1, \ldots, n_s\}$ is linearly independent (so $s < p$). If $c \in \mathbb{R}^s$ and $a \in \mathbb{R}^p$, $Q \in \mathbb{R}^{p \times p}$ are such that $\mathcal{S} = \{x \in \mathbb{R}^p :$

$(x-a)^{\mathrm{T}}Q^{-1}(x-a) \leq 1, {n_k}^{\mathrm{T}}x \leq c^k, k=1,\ldots,s\} \neq \emptyset$, then

$$\begin{aligned}\bar{x} &= a + QN(N^{\mathrm{T}}QN)^{-1}(c - N^{\mathrm{T}}a) \\ &\quad + \sqrt{\frac{1-(c-N^{\mathrm{T}}a)^{\mathrm{T}}(N^{\mathrm{T}}QN)^{-1}(c-N^{\mathrm{T}}a)}{n^{\mathrm{T}}(I-QN(N^{\mathrm{T}}QN)^{-1}N^{\mathrm{T}})Qn}} \\ &\quad \times (I-QN(N^{\mathrm{T}}QN)^{-1}N^{\mathrm{T}})Qn \in \mathcal{S} \qquad \text{and} \qquad (4.19)\end{aligned}$$

$$\bar{x} = \max_{x\in\mathcal{S}} n^{\mathrm{T}}x, \qquad (4.20)$$

$$\begin{aligned}n^{\mathrm{T}}\bar{x} &= n^{\mathrm{T}}a + n^{\mathrm{T}}QN(N^{\mathrm{T}}QN)^{-1}(c-N^{\mathrm{T}}a) \\ &\quad + \sqrt{(1-(c-N^{\mathrm{T}}a)^{\mathrm{T}}(N^{\mathrm{T}}QN)^{-1}(c-N^{\mathrm{T}}a))(n^{\mathrm{T}}(I-QN(N^{\mathrm{T}}QN)^{-1}N^{\mathrm{T}})Qn)}\end{aligned} \qquad (4.21)$$

and

$$\begin{aligned}\hat{x} &= a + QN(N^{\mathrm{T}}QN)^{-1}(c-N^{\mathrm{T}}a) \\ &\quad - \sqrt{\frac{1-(c-N^{\mathrm{T}}a)^{\mathrm{T}}(N^{\mathrm{T}}QN)^{-1}(c-N^{\mathrm{T}}a)}{n^{\mathrm{T}}(I-QN(N^{\mathrm{T}}QN)^{-1}N^{\mathrm{T}})Qn}}(I-QN(N^{\mathrm{T}}QN)^{-1}N^{\mathrm{T}})Qn \in \mathcal{S} \text{ and}\end{aligned}$$

$$\hat{x} = \min_{x\in\mathcal{S}} n^{\mathrm{T}}x,$$

$$\begin{aligned}n^{\mathrm{T}}\hat{x} &= n^{\mathrm{T}}a + n^{\mathrm{T}}QN(N^{\mathrm{T}}QN)^{-1}(c-N^{\mathrm{T}}a) \\ &\quad - \sqrt{(1-(c-N^{\mathrm{T}}a)^{\mathrm{T}}(N^{\mathrm{T}}QN)^{-1}(c-N^{\mathrm{T}}a))(n^{\mathrm{T}}(I-QN(N^{\mathrm{T}}QN)^{-1}N^{\mathrm{T}})Qn)},\end{aligned}$$

where $N = [n_1,\ldots,n_s] \in \mathbb{R}^{p\times s}$. □

Proof 4.1: *P We first note that $N^{\mathrm{T}}QN$ is invertible, as $Q^{1/2}N$ (where $Q^{1/2}$ is any (symmetric) square root of Q) has rank s because $\{n_1,\ldots,n_s\}$ is a linearly independent set and $Q^{1/2}$ is nonsingular. But then $N^{\mathrm{T}}QN$ has full rank s.*

To find $\bar{x}$, we form the Lagrangian

$$\mathcal{L}_0(x,\lambda,\ell) = n^{\mathrm{T}}x + \ell^{\mathrm{T}}(N^{\mathrm{T}}x - c) + \lambda\left((x-a)^{\mathrm{T}}Q^{-1}(x-a) - 1\right),$$

where $\ell \in \mathbb{R}^s$ and λ are formed from Lagrangian multipliers.

If we set $x = a + Q^{1/2}z$, $m_i = Q^{1/2}n_i, i = 1,\ldots,s$, we can use the simpler Lagrangian

$$\mathcal{L}(z,\lambda,\ell) = m^{\mathrm{T}}z + \ell^{\mathrm{T}}(M^{\mathrm{T}}z - d) + \lambda(z^{\mathrm{T}}z - 1),$$

where $m = Q^{1/2}n$, $M = Q^{1/2}N = [Q^{1/2}n_1, \ldots, Q^{1/2}n_s]$, $d = c - N^{\mathrm{T}}a = [c^1 - {n_1}^{\mathrm{T}}a, \ldots, c^s - {n_s}^{\mathrm{T}}a]$ and $\mathcal{L}$ differs from $\mathcal{L}_0$ by a constant.

In line with the usual procedure with Lagrangians, we differentiate this with respect to z^{T}, set the result to zero and solve for z. The resulting z is

$$-\frac{1}{2\lambda}(m + M\ell),$$

and substituting this back in $\mathcal{L}(z, \lambda, \ell)$ yields

$$\mathcal{L}(z(\lambda, \ell), \lambda, \ell) = -\frac{1}{4\lambda}(m + M\ell)^{\mathrm{T}}(m + M\ell) - \ell^{\mathrm{T}}d - \lambda.$$

Then

$$\frac{\partial}{\partial \ell^{\mathrm{T}}}\mathcal{L}(z(\lambda, \ell), \lambda, \ell) = -\frac{1}{2\lambda}(M^{\mathrm{T}}M\ell + M^{\mathrm{T}}m) - d$$

and setting this to zero yields $\ell = -(M^{\mathrm{T}}M)^{-1}(M^{\mathrm{T}}m + 2\lambda d)$

Also,

$$\frac{\partial}{\partial \lambda}\mathcal{L}(z(\lambda, \ell), \lambda, \ell) = -\frac{1}{4\lambda^2}(m + M\ell)^{\mathrm{T}}(m + M\ell) - 1$$

and substituting the above expression for ℓ in this when it is put to zero yields

$$4(d^{\mathrm{T}}(M^{\mathrm{T}}M)^{-1}d - 1)\lambda^2 + m^{\mathrm{T}}(I - M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}})m = 0.$$

Now, if $f \in \mathbb{R}^p$, $\exists f^M \in \mathrm{Im}(M) = \{y \in \mathbb{R}^p : \exists w \in \mathbb{R}^s \text{ such that } y = Mw\}, f^{\perp} \in \mathrm{Im}(M)^{\perp} = \{y \in \mathbb{R}^p : y^{\mathrm{T}}u = 0 \forall u \in \mathrm{Im}(M)\}$ such that $f = f^M + f^{\perp}$. Since $f^{\perp \mathrm{T}}Mu = u^{\mathrm{T}}M^{\mathrm{T}}f^{\perp} = 0 \forall u \in \mathbb{R}^p$, $M^{\mathrm{T}}f^{\perp} = 0$. Let $f^M = Mv$. Then $f^{\mathrm{T}}(I - M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}})f = f^{M\mathrm{T}}f^M + f^{\perp \mathrm{T}}f^{\perp} - v^{\mathrm{T}}M^{\mathrm{T}}M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}Mv = f^{M\mathrm{T}}f^M + f^{\perp \mathrm{T}}f^{\perp} - v^{\mathrm{T}}M^{\mathrm{T}}Mv = f^{\perp \mathrm{T}}f^{\perp} \geq 0$, so $I - M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}$ is positive semi-definite, and positive definite on the subspace $\mathrm{Im}(M)^{\perp}$ and, consequently, it is positive definite on $\mathbb{R}^p - \mathrm{Im}(M)$. But $n \notin \mathrm{Im}(N)$, which implies $m \notin \mathrm{Im}(M)$ Hence, if $1 - d^{\mathrm{T}}(M^{\mathrm{T}}M)^{-1}d > 0$, we obtain

$$\lambda = \pm\frac{1}{2}\sqrt{\frac{m^{\mathrm{T}}(I - M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}})m}{1 - d^{\mathrm{T}}(M^{\mathrm{T}}M)^{-1}d}}$$

and so

$$z = M(M^{\mathrm{T}}M)^{-1}d \mp \sqrt{\frac{1 - d^{\mathrm{T}}(M^{\mathrm{T}}M)^{-1}d}{m^{\mathrm{T}}(I - M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}})m}}(I - M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}})m,$$

and the required expression for $\bar{x}$ follows on setting $\bar{x} = a + Q^{1/2}z$ and choosing $\pm = -$. That this is the appropriate choice for the maximum follows from the comparison of the two possible

expressions for $n^{\mathrm{T}}\bar{x}$. The expression for the value at which the minimum is achieved, $\hat{x}$, is derived by making the other choice of sign. ■

By this Theorem, on replacing n by n_2, N by n and c by $y_1 + 1$, we find that

$$\begin{aligned} x^+_{\mathbb{H}_1^+\varepsilon} &= a - \frac{\nu_1 + 1}{g_1} Q n_1 + \frac{1}{g_1} + \sqrt{\frac{(\nu_1+1)^2 - g_1}{g_1 g_2 - h^2}} (g_1 - Q n_1 n_1{}^{\mathrm{T}}) Q n_2 \\ n_2{}^{\mathrm{T}} x^+_{\mathbb{H}_1^+\varepsilon} &= n_2{}^{\mathrm{T}} a - \frac{\nu_1 + 1}{g_1} h + \frac{\sqrt{((\nu_1+1)^2 - g_1)(g_1 g_2 - h^2)}}{g_1}, \end{aligned}$$

where, of course, $g_i = n_i{}^{\mathrm{T}} Q n_i$, $h = n_1{}^{\mathrm{T}} Q n_2$ and $\nu_1 = y_1 - n_1{}^{\mathrm{T}} a$, and, similarly,

$$\begin{aligned} x^+_{\mathbb{H}_1^-\varepsilon} &= a - \frac{\nu_1 - 1}{g_1} Q n_1 + \frac{1}{g_1} + \sqrt{\frac{(\nu_1-1)^2 - g_1}{g_1 g_2 - h^2}} (g_1 - Q n_1 n_1{}^{\mathrm{T}}) Q n_2 \\ n_2{}^{\mathrm{T}} x^+_{\mathbb{H}_1^-\varepsilon} &= n_2{}^{\mathrm{T}} a - \frac{\nu_1 - 1}{g_1} h + \frac{\sqrt{((\nu_1-1)^2 - g_1)(g_1 g_2 - h^2)}}{g_1}. \end{aligned}$$

Once the expressions for $n_2{}^{\mathrm{T}} x^+_{\mathbb{H}_1^\pm\varepsilon}$ have been found, we can easily find $x^+_{\Pi_1\varepsilon}$ and then shift $\mathbb{H}_2^+$ to pass through it by setting $x^+ = x^+_{\Pi_1\varepsilon}$.

We find x^- in an analogous fashion, and then use $x^\pm$ to find n_2' and y_2'. As $\mathbb{H}_2'^{\pm}$ have to be parallel to $\mathbb{H}_2^\pm$, we have $n_2' = \alpha n_2$. Then $n_2'{}^{\mathrm{T}} x^+ - y_2' = 1$, $n_2'{}^{\mathrm{T}} x^- - y_2' = -1$ lead to $\alpha = 2/n_2{}^{\mathrm{T}}(x^+ - x^-)$, $y_2' = n_2{}^{\mathrm{T}}(x^+ + x^-)/n_2{}^{\mathrm{T}}(x^+ - x^-)$.

4.2.2 Which Hyperplane Pairs?

The simplest choice is to take successive pairs, and this is done for the first variation on the unmodified Fogel-Huang algorithm; that is, the unmodified Fogel-Huang algorithm is performed on the initial ellipsoid and the first hyperplane encountered, then the modified two pair variant is performed on the resulting ellipsoid and the first and second hyperplane pairs to produce the third ellipsoid, and then the second and third hyperplane pairs are utilised to derive the fourth ellipsoid, to that the kth ellipsoid is generated from the $(k-1)$st ellipsoid and the $(k-2)$nd and $(k-1)$st hyperplane pairs. This will be called the *strict sequence* variant.

The second variant, the *odd/even sequence* variant, takes account of the fact that in the strict sequence variant, a hyperplane pair is utilised in the production of two successive ellipsoids, and the first of these ellipsoids will already contain, in some sense, "information" from that hyperplane pair before the hyperplane pair is used in the generation of the second of the ellipsoids. The sequence of ellipsoids in this odd/even sequence variant will be: first, in the unmodified Fogel-Huang; first and second; first and third; second and fourth; third and fifth; and so on, in the two pair variant.

The final variant considered here for a single pass through the data, is to take the current hyperplane and try to predict which of the hyperplane pairs already encountered (possibly as shifted *á la* Belforte, Bona and Cerone) will result in the smallest ellipsoid when used with the current hyperplane in the two pair variant on the Fogel-Huang algorithm.

If the current pair is the kth, $j \in \{1, \ldots, k-1\}$ such that

$$\left.\frac{\partial}{\partial q_j} \det \bar{Q}(q_j, q_k)\right|_{(q_1,q_2)=0}$$

is minimised is chosen and the jth hyperplane pair is used (where $\mathcal{E}(\bar{a}(q_i, q_j), \bar{Q}(q_i, q_j))$ is the family of ellipsoids considered in Section 4.1 — under $q_1 \to q_j$, $q_2 \to q_k$), where Q is the shape matrix of the current ellipsoid. In effect, the choice is of the hyperplane pair which most sharply reduces $\det \bar{Q}$ when the corresponding q is increased from zero.

But

$$\left.\frac{\partial}{\partial q_j} \det \bar{Q}(q_j, q_k)\right|_{(q_1,q_2)=0} = \det Q[p(1-\nu_j^2) - g_j],$$

so minimising $p(1-\nu_j^2) - g_j$ produces the desired ellipsoid.

Of course, when this *best second pair* variant is started off, there is only one hyperplane pair and this is used in the unmodified Fogel-Huang algorithm.

4.3 s Pairs at a Time

If our two pairs of hyperplanes have normals which are orthogonal in the metric of our ellipsoid, i.e., $h = {n_1}^{\mathrm{T}} Q n_2 = 0$, then equations (4.6), (4.5) and (4.8) specialise to

$$\begin{aligned}
\bar{Q}(q_1, q_2) &= Q\left[I_p - \frac{q_1}{1+g_1q_1} n_1 {n_1}^{\mathrm{T}} Q - \frac{q_2}{1+g_2q_2} n_2 {n_2}^{\mathrm{T}} Q\right], \\
\bar{a}(q_1, q_2) &= a + Q\left[\frac{\nu_1 q_1}{1+g_1q_1} n_1 + \frac{\nu_2 q_2}{1+g_2q_2} n_2\right]
\end{aligned}$$

and

$$\tilde{Q}(q_1, q_2) = \left[1 + q_1 + q_2 - \frac{\nu_1^2 q_1}{1+g_1q_1} - \frac{\nu_2^2 q_2}{1+g_2q_2}\right] \bar{Q}(q_1, q_2),$$

so equation (4.9) becomes

$$\det \tilde{Q}(q_1, q_2) = \frac{\det Q}{(1+g_1q_1)(1+g_2q_2)} \left[1 + q_1 + q_2 - \frac{\nu_1^2 q_1}{1+g_1q_1} - \frac{\nu_2^2 q_2}{1+g_2q_2}\right]^p.$$

If we intersect s Q-orthogonal hyperplane pairs Π_i with our ellipsoid $\mathcal{E}(a, Q)$, we can add q_i times the inequality defining each Π_i to the inequality defining $\mathcal{E}(a, Q)$, to obtain an inequality

corresponding to (4.3). Since this last inequality will have as its left-hand side a positive-definite quadratic form in $x-\bar{a}$, for some $\bar{a}$, it too will define an ellipsoid, $\mathcal{E}(\bar{a}(q_1,\ldots,q_s),\bar{Q}(q_1,\ldots,q_s)) = \mathcal{E}(\bar{a}(q),\bar{Q}(q))$ where now $q=(q_1,\ldots,q_s)$. $\bar{a}$ and $\bar{Q}$ defining this ellipsoid, together with $\det\bar{Q}$, are given by the following equations, derived by induction from the preceding ones:

$$\bar{a}(q_1,\ldots,q_s) = a + Q\sum_{k=1}^{s}\frac{\nu_k q_k}{1+g_k q_k}n_k, \tag{4.22}$$

$$\tilde{Q}(q_1,\ldots,q_s) = Q - Q\sum_{k=1}^{s}\frac{q_k}{1+g_k q_k}n_k {n_k}^{\mathrm{T}} Q, \tag{4.23}$$

$$\bar{Q}(q_1,\ldots,q_s) = \left[1+\sum_{k=1}^{s}\left(q_k - \frac{\nu_k^2 q_k}{1+g_k q_k}\right)\right]\tilde{Q}(q_1,\ldots,q_s), \tag{4.24}$$

$$\det\bar{Q}(q_1,\ldots,q_s) = \left[1+\sum_{k=1}^{s}\left(q_k - \frac{\nu_k^2 q_k}{1+g_k q_k}\right)\right]^{p}\frac{\det Q}{\prod_{k=1}^{s}(1+g_k q_k)}, \tag{4.25}$$

where $g_i = {n_i}^{\mathrm{T}} Q n_i$, $\nu_i = y_i - {n_i}^{\mathrm{T}} a$.

We note that the single relative minimum of $q_k - \nu_k^2 q_k/(1+g_k q_k)$ on $(-g_k^{-1},\infty)$ is at $q_k = (|\nu_k|-1)/g_k$ and that this value is $-(1-|\nu_k|)^2/g_k$. Hence, the minimum value of $q_k - \nu_k^2 q_k/(1+g_k q_k)$ on $[0,\infty)$ is 0, if $|\nu_k| \le 1$ and $q_k - \nu_k^2 q_k/(1+g_k q_k)$, if $\nu_k| \ge 1$. Then

$$1 - \sum_{1\le k\le s, |\nu_k|>1}\frac{(1-|\nu_k|)^2}{g_k} < 0$$

implies $\exists q \in O_+ := \{q : q_k \ge 0\}$ such that $\det\bar{Q}(q) < 0$, which would mean that $\mathcal{E}(a,Q)\cap\Pi_1\cap\ldots\cap\Pi_s = \emptyset$.

Hence, we only consider hyperplane pairs $\mathbb{H}_1^{\pm},\ldots,\mathbb{H}_s^{\pm}$ such that

$$1 - \sum_{1\le k\le s, |\nu_k|>1}\frac{(1-|\nu_k|)^2}{g_k} \ge 0. \tag{4.26}$$

In order to find the infimum of $\det\bar{Q}(q)$, which is proportional to the square of the volume of $\mathcal{E}(\bar{a}(q),\bar{Q}(q))$, over O_+, we set

$$\Lambda(q) = \frac{\det\bar{Q}(q)}{\det Q} \tag{4.27}$$

We first examine what happens if $q_i \to \infty$ for any i such that $\le i \le s$. As $q_i\left(1-\frac{\nu_i^2}{1+g_i q_i}\right) \to \infty$ as $q_i \to \infty$, $1+\sum_{j=1}^{s}\left(1-\frac{\nu_j^2}{1+g_j q_j}\right) \to \infty$ as $q_i \to \infty$. Thus, equation (4.25) means that $\Lambda \to \infty$ as $q_i \to \infty$ unless $s=p$ and $q_j \to \infty$, $1 \le j \le p$. As we are looking for the infimum, we concentrate on this exceptional case. Then, $\Lambda(q)$ can be written

$$\Lambda(q) = \prod_{k=1}^{p}\left[\frac{1}{1+g_k q_k} + \sum_{l=1}^{p}\frac{q_l}{1+g_k q_k}\left(1-\frac{\nu_l^2}{1+g_l q_l}\right)\right],$$

so, if any $q_k/q_l \to \infty$ as the $q_j \to \infty$, $\Lambda(q) \to \infty$, and hence, if $\Lambda(q)$ is to remain bounded, we must have $q \to (r_1, \ldots, r_p)r$ as $q_i \to \infty$, where $r \to \infty$ and each $r_j \in (0, \infty)$. Then

$$\Lambda(q) \to \prod_{k=1}^{p} \left[\sum_{l=1}^{p} \frac{r_l}{g_k r_k} \right] = \frac{\left(\sum_{k=1}^{p} r_k\right)^p}{\prod_{k=1}^{p} g_k r_k},$$

and we now wish to minimise this quantity. As it is homogeneous in $(r_1, \ldots, r_p)$, we can set $r_1 = 1$. Then differentiation with respect to r_i, $i = 2, \ldots, p$ and setting each of the results to zero reveals that $r_2 = \cdots = r_p = 1$, so

$$\Lambda(q) \to \frac{p^p}{\prod_{k=1}^{p} g_k}, \text{ as } q \to \infty \text{ along } q_1 = q_2 = \cdots = q_p, \tag{4.28}$$

and this is the smallest value approached by $\Lambda(q)$ as any $q_i \to \infty$.

We now look at the behaviour of $\tilde{Q}(q)$, $\bar{Q}(q)$ and $\bar{a}(q)$ as q becomes unbounded along $q_1 = q_2 = \ldots = q_p = r$, say. Then

$$\begin{aligned} \tilde{Q}(q) &= Q \left[\mathrm{I}_p - \sum_{k=1}^{p} \frac{r}{1 + g_k r} n_k n_k^{\mathrm{T}} Q \right] \\ &= Q \left[\mathrm{I}_p - \sum_{k=1}^{p} \frac{1}{g_k} n_k n_k^{\mathrm{T}} Q + \frac{1}{r} \sum_{k=1}^{p} \frac{1}{g_k^2} n_k n_k^{\mathrm{T}} Q \right] + \mathrm{O}(r^{-2}). \end{aligned}$$

But $\left(\sum_{k=1}^{p} \frac{1}{g_k} n_k n_k^{\mathrm{T}} Q\right) n_i = \frac{1}{g_i} \left(n_i^{\mathrm{T}} Q n_i\right) n_i = n_i$, $i = 1, \ldots, p$, so we must have $\sum_{k=1}^{p} \frac{1}{g_k} n_k n_k^{\mathrm{T}} Q = I_p$ and

$$\tilde{Q}(q) = \frac{1}{r} Q \sum_{k=1}^{p} g_k^{-2} n_k n_k^{\mathrm{T}} Q + \mathrm{O}(r^{-2}),$$

with the consequence that

$$\begin{aligned} \bar{Q}(q) &= \left[1 + r \sum_{k=1}^{p} \left(1 - \frac{\nu_k^2}{1 + g_k r} \right) \right] \tilde{Q}(q) \\ &= \left(pr + \mathrm{O}(r^0) \right) \tilde{Q}(q) \\ &\to pQ \sum_{k=1}^{p} g_k^{-2} n_k n_k^{\mathrm{T}} Q. \end{aligned}$$

Also,

$$\begin{aligned} \bar{a}(q) &= a + Q \sum_{k=1}^{p} \frac{\nu_k r}{1 + g_k r} n_k \\ &= a + Q \sum_{k=1}^{p} \frac{\nu_k}{g_k} n_k + \mathrm{O}(r^{-1}) \\ &\to \left[I_p - Q \sum_{k=1}^{p} g_k^{-1} n_k n_k^{\mathrm{T}} \right] a + \sum_{k=1}^{p} y_k g_k^{-1} Q n_k \\ &= \sum_{k=1}^{p} y_k g_k^{-1} Q n_k. \end{aligned}$$

Now consider the parallelotope $\Pi_1 \cap \ldots \cap \Pi_p$. Its 2^p vertices x_u are given by

$$n_i^{\mathrm{T}} x_u = y_i + u_i,$$

where $u_i \in \{-1, 1\}$, $i = 1, \ldots, p$.

Writing $N = [n_1, \ldots, n_p]^{\mathrm{T}}$, $y = [y_1, \ldots, y_p]^{\mathrm{T}}$, we have

$$Nx_u = y + u, \quad \forall u \in \mathbb{Z}_2^p, \mathbb{Z}_2 = \{-1, 1\}.$$

Since the n_i satisfy $n_i \neq 0$, $n_i^{\mathrm{T}} Q n_j = 0$, $i \neq j$, and Q is positive definite, they are linearly independent and N is invertible. We seek an affine co-ordinate transformation $z = P^{-1}x + z_0$ such that $x_u = P(u - z_0)$. Then $Nx_u = NPu - NPz_0 = u + y$. Clearly, $P = N^{-1}$, $z_0 = -y$ fits the bill. In the transformed co-ordinates, the image of $\Pi_1 \cap \ldots \cap \Pi_p$ is just the cube with vertices in $\mathbb{Z}_2^p$, and the minimum-volume ellipsoid containing this cube is just $\mathcal{E}(0, pI_p)$. As the ordering of volumes is preserved by affine transformations, the minimum-volume ellipsoid containing $\Pi_1 \cap \ldots \cap \Pi_p$ is the image of $\mathcal{E}(0, pI_p)$ under the inverse transformation: $\mathcal{E}(N^{-1}y, pN^{-1}N^{-\mathrm{T}})$. But $NQN^{\mathrm{T}} = \mathrm{diag}(g_1, \ldots, g_p) := G$, so $N^{-1} = QN^{\mathrm{T}}G^{-1}$ and then $N^{-1}y = Q[n_1, \ldots, n_p] \times G^{-1}(y_1, \ldots, y_p)^{\mathrm{T}} = \sum_{k=1}^p \frac{y_k}{g_k} Q n_k$, and $pN^{-1}N^{-\mathrm{T}} = pQN^{\mathrm{T}}G^{-2}NQ = pQ\left(\sum_{k=1}^p g_k^{-2} n_k n_k^{\mathrm{T}}\right)Q$. Thus the ellipsoid approached by $\mathcal{E}(\bar{a}(q), \bar{Q}(q))$ when $q = (1, \ldots, 1)r$ and $r \to \infty$ is the minimum volume ellipsoid containing $\Pi_1 \cap \ldots \cap \Pi_p$.

At this point we make a definition:

Definition 4.1: *We set*

$$\Lambda_\infty = \frac{p^p}{\prod_{k=1}^p g_k}.$$

□

If $\Lambda_\infty > \inf_{q \in O_+}\{\Lambda(q)\}$, then this infimum is attained at a point of a compact subset of O_+. Bearing this possibility in mind, we seek $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_s)$ such that

$$\left.\frac{\partial \Lambda}{\partial q_k}\right|_{q=\hat{q}} = 0, \; k \in K \subset \{1, \ldots, s\}, q_k = 0, k \in \{1, \ldots, s\} - K \tag{4.29}$$

i.e.

$$\Lambda \left[p\left(1 - \frac{\nu_k^2}{(1 + g_k \hat{q}_k)^2}\right) \left\{ 1 + \sum_{l \in K} \left(\hat{q}_l - \frac{\nu_l^2 \hat{q}_l}{1 + g_l \hat{q}_l} \right) \right\}^{-1} - \frac{g_k}{1 + g_k \hat{q}_k} \right] = 0, k \in K, \tag{4.30}$$

because a necessary condition for a minimum on a set is that all the first derivatives vanish at $\hat{q}$ *or* that $\hat{q} \in \partial O_+$, the boundary of O_+ (given by $q_k = 0$ for some k). But this argument can also

be applied to ∂O_+. As $\Lambda > 0$ (otherwise the resultant ellipsoid is degenerate), equation (4.30) can be written

$$\frac{1}{g_k}\left(1 + g_k\hat{q}_k - \frac{\nu_k^2}{1 + g_k\hat{q}_k}\right) = \frac{1}{p}\left[1 + \sum_{l\in K}\left(\hat{q}_l - \frac{\nu_l^2\hat{q}_l}{1 + g_l\hat{q}_l}\right)\right] \tag{4.31}$$

where the right-hand side is independent of k. If we set this right-hand side equal to A, we have

$$A = \frac{1}{p}\left[1 + \sum_{l\in K}\left(\hat{q}_l - \frac{\nu_l^2\hat{q}_l}{1 + g_l\hat{q}_l}\right)\right] \tag{4.32}$$

which, by the reasoning leading to inequality (4.26), means

$$A \geq \frac{1}{p}\left[1 - \sum_{1\leq k\leq s, |\nu_k|>1}\frac{(1-|\nu_k|)^2}{g_k}\right] \geq 0, \quad A \geq \frac{1}{p}\left[1 - \sum_{1\leq k\leq s,}\frac{(1-|\nu_k|)^2}{g_k}\right]. \tag{4.33}$$

Also, equation (4.31) can be expressed as the quadratic

$$(1 + g_k\hat{q}_k)^2 - g_kA(1 + g_k\hat{q}_k) - \nu_k^2 = 0, \tag{4.34}$$

leading to

$$\begin{aligned} 1 + g_k\hat{q}_k &= \frac{1}{2}\left[Ag_k \pm \sqrt{A^2g_k^2 + 4\nu_k^2}\right] \\ &= \frac{1}{2}\left[Ag_k + \sqrt{A^2g_k^2 + 4\nu_k^2}\right], \end{aligned} \tag{4.35}$$

as the left-hand side is nonnegative, and

$$\hat{q}_k = \frac{1}{2g_k}\left[Ag_k - 2 + \sqrt{A^2g_k^2 + 4\nu_k^2}\right]. \tag{4.36}$$

As $\hat{q}_k \geq 0 \forall k \in K$, we can put a condition on A: either $A \geq 2/g_k$ or $(Ag_k - 2)^2 \leq A^2g_k^2 + 4\nu_k^2 \Rightarrow$ $A \geq (1 - \nu_k^2)/g_k$. As this second bound is implied by the first, we have the condition

$$A \geq \max_{k\in K}\left[\frac{1-\nu_k^2}{g_k}\right], \tag{4.37}$$

and the condition (4.33) will be a consequence of this.

We can now substitute equation (4.36) in the definition of A, equation (4.32), to obtain an equation not involving $\hat{q}$ for A:

$$A = \frac{1}{p}\left[1 - \sum_{k\in K}\frac{1+\nu_k^2}{g_k} + \sum_{k\in K}\left(A^2 + \frac{4\nu_k^2}{g_k^2}\right)^{1/2}\right]. \tag{4.38}$$

To enable us to analyse this equation for various subsets K of $\mathbb{N}_s := \{1, \ldots, s\}$, we make some definitions:

Definition 4.2: *For each subset K, we let*

$$\begin{aligned} a_K &:= \frac{1}{p}\left(1 - \sum_{k\in K}\frac{1+\nu_k^2}{g_k}\right), \\ b_K &:= \frac{\overline{\overline{K}}}{p}, \\ c_K &:= \frac{2}{p}\sum_{k\in K}\frac{\nu_k^2}{g_k^2}, \\ d_K &:= \frac{2}{p}\sum_{k\in K}\frac{|\nu_k|}{g_k}, \end{aligned}$$

where $\overline{\overline{S}}$ is the cardinality of a set S. □

We will assume that at least one hyperplane from each hyperplane pair will intersect or, at worst, touch our ellipsoid (if necessary, we will move them to positions parallel to their old placing), so we may assume that (see Subsection 3.1.1)

$$|\nu_k| \leq \sqrt{g_k} - 1,\ g \geq 1\ \forall k \in \mathbb{N}_s. \tag{4.39}$$

Inequalities (4.39) enable us to make deductions about the quantities in Definition 4.2. For example, the maximum values of $(1+\nu^2)/g$, $(1-|\nu|)^2/g$ and $|\nu|/g$ on the set $\{(g,\nu) : g \geq 1, |\nu| \leq \sqrt{g} - 1\}$ are 1, 1 and $\frac{1}{4}$, respectively (at $(g,\nu) = (1,0)$, $(g,\nu) = (1,0)$ and $(g,\nu) = (4,\pm 1)$, respectively), and the infima of these quantities on this set are 0, so

$$\begin{array}{rcccl} p^{-1} - b_K = -\frac{\overline{\overline{K}}-1}{p} & \leq & a_K & \leq & p^{-1}, \\ -\frac{\overline{\overline{K}}-1}{p} & \leq & a_K + d_K & \leq & p^{-1}, \\ 0 & \leq & c_K & \leq & \frac{1}{8}b_K, \\ 0 & \leq & d_K & \leq & \frac{1}{2}b_K. \end{array} \tag{4.40}$$

We can now make a further definition:

Definition 4.3:

$$F_K : \mathbb{R} \to \mathbb{R} : F_K(A) = a_K + \frac{1}{p}\left[\sum_{k\in K}\left(A^2 + \frac{4\nu_k^2}{g_k^2}\right)^{1/2}\right].$$

□

Theorem 4.2: The equation $F_K(A) = A$ has a unique solution except when $\overline{\overline{K}} = p$ and

a) $\nu_k = 0\ \forall k \in K$ and $\sum_{k\in K} g_k^{-1} = 1$, when $F_K(A) = |A|$ and every nonnegative A is a solution;

or b) $\nu_k = 0\ \forall k \in K$ and $\sum_{k\in K} g_k^{-1} < 1$, when $F_K(A) = |A| + \frac{1}{p}\left(1 - \sum_{k=1}^{p}\right) > A$, so the equation has no solution;

or c) $\exists k \neq 0$ such that $\nu_k \neq 0$ and $a_K \geq 0$, when the equation has no solution.

Moreover, if $F_K(A) = A$ has a unique solution, it has the same sign as $a_K + d_K$. □

Proof 4.2: *P The proofs for a) and b) are contained in their statements.*

In the proof for c), we suppose $\overline{\overline{K}} = p$ *and* $a_K \geq 0$. *Then, if* $A \in \mathbb{R}$,

$$\begin{aligned} F_K(A) &= a_K + \frac{1}{p}\left[\sum_{k\in K}\left(A^2 + \frac{4\nu_k^2}{g_k^2}\right)^{1/2}\right] \\ &> \frac{1}{p}a_K + |A| \geq A, \end{aligned}$$

so $F_K(A) > A \forall A$ *in this case.*

We start by proving the existence of fixed points of F_K.

$$\begin{aligned} F_K(-|a_K|) &= a_K + \frac{1}{p}\sum_{k\in K}\left(a_K^2 + \frac{4\nu_k^2}{g_k^2}\right)^{1/2} \\ &\geq a_K + b_K|a_K| \geq -|a_K|. \end{aligned}$$

Also, if $a_K < 0$,

$$\begin{aligned} F_K\left(\frac{c_K}{|a_K|}\right) &= \frac{1}{p}\sum_{k\in K}\left(\frac{c_K^2}{a_K^2} + \frac{4\nu_k^2}{g_k^2}\right)^{1/2} \\ &\leq a_K + \frac{1}{p}\sum_{k\in K}\left(\frac{c_K}{|a_K|} + \frac{2|a_K|}{c_K}\,\frac{\nu_k^2}{g_k^2}\right) \\ &= a_K + b_K\frac{c_K}{|a_K|} + |a_K| \\ &= b_K\,\frac{c_K}{|a_K|} \leq \frac{c_K}{|a_K|}, \end{aligned}$$

and, if $a_K \geq 0$, $b_K < 1$

$$\begin{aligned} F_K\left(\frac{a_K + d_K}{1 - b_K}\right) &= a_K + \frac{1}{p}\sum_{k\in K}\left(\left(\frac{a_K + d_K}{1 - b_K}\right)^2 + \frac{4\nu_k^2}{g_k^2}\right)^{1/2} \\ &\leq a_K + \frac{1}{p}\sum_{k\in K}\left(\frac{a_K + d_K}{1 - b_K} + \frac{2|\nu_k|}{g_k}\right) \\ &= a_K + \frac{b_K(a_K + d_K)}{1 - b_K} + d_K \\ &= \frac{a_K + d_K}{1 - b_K}. \end{aligned}$$

Hence, if either $a_K < 0$, *or* $a_K \geq 0$ *and* $b_K < 1$, $\exists A_-$ *and* $A_+ \in \mathbb{R}$ *such that* $F_K(A_-) \geq A_-$ *and* $F_K(A_+) \leq A_+$. *We assume that neither* A_- *nor* A_+ *are themselves fixed points of* F_K. *Then, if* $G_K = F_K - \iota$, *where* ι *is the identity function,* $G_K(A_-) > 0$ *and* $G_K(A_+) < 0$. *But* F_K, *and hence* G_K, *is a continuous function, so, by the mean-value theorem of real analysis,* $\exists A \in \mathbb{R}$ *such that* $G_K(A) = 0$ *or* $F_K(A) = A$.

We prove the uniqueness of the fixed point by looking at the derivative of G_K *(which exists everywhere except at* 0, *and it exists there too if* $\nu_k \neq 0 \forall k$*). Now*

$$\begin{aligned} \frac{\mathrm{d}G_K}{\mathrm{d}A} &= \frac{1}{p}\sum_{k\in K} A\left(A^2 + \frac{4\nu_k^2}{g_k^2}\right)^{-1/2} - 1 \\ &\leq b_K - 1 \\ &\leq 0, \forall A \in \mathbb{R} - \{0\} \text{ (and also for } A = 0 \text{ if } \nu_k \neq 0 \forall k \in K) \end{aligned}$$

where the first equality holds only if $\nu_k = 0 \forall k \in K$ *and the second only if* $\overline{\overline{K}} = p$. *Hence,* $\mathrm{d}G_K/\mathrm{d}A < 0$ *when a) in the statement of the theorem does not hold, and so* G_K *is strictly decreasing (the possible failure of* $\mathrm{d}G_K/\mathrm{d}A$ *to be defined at* $A = 0$ *does not affect this conclusion). Hence,* $F_K > \iota$ *to the left of any fixed point of* F_K *and* $F_K < \iota$ *to the right of that fixed point, so there is at most one such fixed point.*

Suppose

$$1 - \sum_{k\in K} \frac{1+\nu_k^2}{g_k^2} = p(a_K + d_K) = 0.$$

Then $F_K(0) = a_K + d_K = 0$, *so* 0 *is a fixed point of* F_K.

Suppose $a_K + d_K > 0$. *Then* $F_K(A) > a_K + d_K > 0$, *so any fixed point of* F_K *must be positive if* $a_K + d_K$ *is.*

Finally, suppose $a_K + d_K < 0$. *Then, if* $A \geq 0$, $F_K(A) \leq a_K + b_K A + d_K \leq a_K + d_K + A < A$, *a contradiction, so* A *has the same sign as* $a_K + d_K$ *here too.* ■

Thus, we can make a further definition:

Definition 4.4: *If*

a) $s < p$ *and* $K \subset \mathbb{N}_s$;

or b) $s = p$ *and* $K \overset{\subset}{\neq} \mathbb{N}_p$;

or c) $s = p$, $K = \mathbb{N}_p$ *and* $a_K \leq 0$,

we define A_K *to be the unique* A *such that* $F_K(A) = A$. □

Now we can relate A_K to other values of A:

Theorem 4.3: If K is such that A_K is defined, then $A > F_K(A)$ (respectively, $A = F_K(A)$, $A < F_K(A)$) is equivalent to $A > A_K$ (respectively, $A = A_K$, $A < A_K$). □

Proof 4.3: *P Simple consequence of the facts that $G_K = F_K - \iota$ is a decreasing function and $G_K(A_K) = 0$.* ■

We first need a lemma to enable us to justify a method for finding A_K:

Lemma 4.4: If $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with constant L, i.e.,

$$|f(x) - f(y)| \leq L|x - y| \; \forall x, y \in \mathbb{R},$$

and $L < 1$, the sequence $\{x_n\}$ defined by $x_1 \in \mathbb{R}$, $x_{n+1} = f(x_n)$ converges to a unique x (dependent only on x_1). The rate of convergence is given by

$$|x - x_n| \leq \alpha L^{n-1},$$

where

$$\alpha = \frac{|x_2 - x_1|}{1 - L}.$$

□

Proof 4.4: *L We have*

$$\begin{aligned} |x_{n+1} - x_n| &= |f(x_n) - f(x_{n-1})| \\ &\leq L|x_n - x_{n-1}| \\ &\leq L^{n-1}|x_2 - x_1|, \end{aligned}$$

so

$$\begin{aligned} |x_{n+k} - x_n| &\leq |x_{n+k} - x_{n+k-1}| + \ldots + |x_{n+1} - x_n| \\ &\leq \left(L^{n+k-2} + L^{n+k-3} + \ldots + L^{n-1}\right) |x_2 - x_1| \\ &= L^{n-1} \frac{1 - L^k}{1 - L} |x_2 - x_1| \end{aligned}$$

and $\{x_n\}$ is clearly a Cauchy sequence. This also means that

$$|x_n - x_1| \leq \frac{1 - L^{n-1}}{1 - L} |x_2 - x_1| < \frac{|x_2 - x_1|}{1 - L}$$

or

$$x_1 - \frac{|x_2 - x_1|}{1 - L} < x_n < x_1 + \frac{|x_2 - x_1|}{1 - L},$$

so $\{x_n\}$ is a sequence in the closed, bounded subset $[x_1 - |x_2 - x_1|/(1 - L), x_1 + |x_2 - x_1|/(1 - L)]$ of $\mathbb{R}$ and so has a limit point (see Goffman [9]), x, say. But a limit point of a Cauchy sequence is the limit of that sequence. Hence, $\{x_n\}$ converges to x.

Letting $k \to \infty$ in the above equation involving it, we also find

$$|x - x_n| \leq \frac{L^{n-1}}{1-L} |x_2 - x_1|$$

■

We also have the following:

Theorem 4.5: If A_K is defined, the sequence $\{A_K^{(n)}\}$ defined by $A_K^{(n)} = F_K(A_K^{(n-1)})$, where

$$A_K^{(0)} = \begin{cases} 0, & a_K < 0; \\ \frac{a_K}{1-b_K}, & a_K \geq 0, b_K < 1 \end{cases}$$

converges to A_K, and

$$\left| A_K - A_K^{(n)} \right| \leq L_K^{n-1} \alpha_K$$

where $L_K \in (0,1)$ is given by

$$L_K = \frac{1}{p} \sum_{k \in K} \beta \left(\beta^2 + \frac{4\nu_k^2}{g_k^2} \right)^{-1/2}$$

for

$$\beta = \begin{cases} \max \left\{ -a_K, \frac{-c_K}{a_K} \right\}, & a_K < 0; \\ \frac{a_K + d_K}{1-b_K}, & a_K > 0, b_K < 1 \end{cases}$$

and

$$\alpha_K = \begin{cases} \frac{|a_K + d_K|}{1-L_K}, & a_K \leq 0; \\ \frac{d_K}{1-L_K}, & a_K > 0. \end{cases}$$

□

Proof 4.5: *P Consider the function $f : \mathbb{R}_+ \to \mathbb{R}_+ : x \mapsto \sqrt{x^2 + a^2}$ defined on the positive reals (where $a \in \mathbb{R}_+$). If $y > x$, it is a matter of simple algebra to show that*

$$\frac{x}{\sqrt{x^2 + a^2}} \leq \frac{f(y) - f(x)}{y - x} \leq \frac{y}{\sqrt{y^2 + a^2}},$$

or, more generally,

$$\min \left\{ \frac{x}{\sqrt{x^2 + a^2}}, \frac{y}{\sqrt{y^2 + a^2}} \right\} \leq \frac{|f(y) - f(x)|}{|y - x|} \leq \max \left\{ \frac{x}{\sqrt{x^2 + a^2}}, \frac{y}{\sqrt{y^2 + a^2}} \right\},$$

for $x \neq y$.

Consequently,

$$\begin{aligned} |F_K(A) - F_K(A')| &\leq \max \left\{ \frac{1}{p} \sum_{k \in K} \frac{|A|}{\sqrt{A^2 + \frac{4\nu_k^2}{g_k^2}}}, \frac{1}{p} \sum_{k \in K} \frac{|A'|}{\sqrt{A'^2 + \frac{4\nu_k^2}{g_k^2}}} \right\} ||A| - |A'|| \\ &\leq \max \left\{ \sum_{k \in K} \frac{|A|}{\sqrt{A^2 + \frac{4\nu_k^2}{g_k^2}}}, \sum_{k \in K} \frac{|A'|}{\sqrt{A'^2 + \frac{4\nu_k^2}{g_k^2}}} \right\} |A - A'| \end{aligned}$$

By the working of Theorem 4.2, $F_K([\beta_1,\beta_2]) \subset [\beta_1,\beta_2]$*, where*

$$[\beta_1,\beta_2] = \begin{cases} [a_K, -c_K/a_K], & a_K < 0 \\ \left[-a_K, \frac{a_K+d_K}{1-b_K}\right] & a_K \geq 0, b_K < 1, \end{cases}$$

so the restriction of F_K *to* $[\beta_1,\beta_2]$ *is a Lipschitz continuous function with constant* L_K *given by*

$$\frac{1}{p}\sum_{k\in K} \frac{\beta}{\sqrt{\beta^2 + \frac{4\nu_k^2}{g_k^2}}},\ \beta = \max\{|\beta_1|, |\beta_2|\}.$$

As this L_K *is clearly less than 1 (in all cases where* A_K *is defined), the restriction of* F_K *to* $[\beta_1,\beta_2]$ *obeys Lemma 4.4 (suitably modified to comply with the restriction), and, as* $A_K^{(n)}$ *lies in* $[\beta_1,\beta_2]$ *if* $A_K^{(0)}$ *does (by a simple inductive argument), the theorem holds for some value of* α_K*, because the values of* $A_K^{(0)}$ *given above do lie in the appropriate intervals, and the value of* β *given there is* $\max\{|\beta_1|, |\beta_2|\}$ *in each case.*

Finally, we show that the correct α*'s are given in the statement of the theorem. If* $a_K \leq 0$ *and* $A_K^{(0)} = 0$*, then* $|A_K^{(0)} - A_K^{(1)}| = |a_K + d_K|$*, as required.*

As $|x| \leq \sqrt{x^2+a^2} \leq |x| + |a|$*,*

$$A_K^{(0)} - a_K - b_K|A_K^{(0)}| \geq A_K^{(0)} - A_K^{(1)} \geq A_K^{(0)} - a_K - b_K|A_K^{(0)}| - d_K,$$

so if $0 < a_K \leq a_K + d_K$*,* $b_K < 1$ *and* $A_K^{(0)} = a_K/(1-b_K) > 0$*, we have*

$$(1-b_K)A_K^{(0)} - a_K = 0 \geq A_K^{(0)} - A_K^{(1)} \geq (1-b_K)A_K^{(0)} - a_K - d_K = -d_K,$$

so $|A_K^{(0)} - A_K^{(1)}| \leq d_K$ *which is a better bound than* $|a_K + d_K|$ *in this case.* ■

We need a few more definitions to allow us to find the smallest of the volumes given by minimising $\Lambda(q)$.

Definition 4.5: *The solution* A_K *of* $F_K(A) = A$*, if defined, is called* valid *if* $A_K \geq 0$ *and* $A_K \geq \max_{k\in K}(1-\nu_k^2)/g_k$*. Otherwise, it is* invalid. □

This means that a valid A_K obeys inequalities (4.37) and (4.33).

Definition 4.6: *If* A_K *is valid, let* $q_K \in (O_+)^p$ *be an arbitrary member of the set* $\{q \in (O_+)^p : q_i = 0, i \notin K\}$*, and let* $\hat{q}_K$ *be the member of the same set given by*

$$\hat{q}_k = \frac{1}{2g_k}\left[g_k A_K - 2 + \sqrt{g_k^2 A_K^2 + 4\nu_k^2}\right].$$

□

Definition 4.7: *Consider* $\Lambda(q)$ *as defined by equations (4.27) and (4.25). At* $q = \hat{q}_K$*, where* A_K *is valid, we have*

$$\Lambda(\hat{q}_K) = \frac{p^p A_K^p}{\prod_{k\in K} \frac{1}{2}\left[g_k A_K + \sqrt{g_k^2 A_K^2 + 4\nu_k^2}\right]},$$

by the definition of A, equation (4.32), the Definitions 4.3 and 4.4 and equation (4.35). By analogy to this, we now define a function $\Lambda_K : \mathbb{R} \to \mathbb{R}$ *by*[1]

$$\Lambda_K(A) = \frac{p^p A^p}{\prod_{k\in K} \frac{1}{2}\left[g_k A + \sqrt{g_k^2 A^2 + 4\nu_k^2}\right]}.$$

□

Definition 4.8: *Let* $k_1 \in K \subset \mathbb{N}_s$, $K - \{k_1\} = J \neq \emptyset$. *We define* $G_{K,k_1} : \mathbb{R} \times [0,1] \to \mathbb{R}$ *by*

$$G_{K,k_1}(D,\lambda) = \frac{1}{p}\left[1 - \sum_{k\in K}\frac{1+\nu_k^2}{g_k} + \sum_{k\in K}\left(D^2 + \frac{4\nu_k^2}{g_k^2}\right)^{1/2} - \lambda\left\{\left(D^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}\right)^{1/2} - \frac{1+\nu_{k_1}^2}{g_{k_1}}\right\}\right].$$

Then $G_{K,k_1}(D,0) = F_K(D)$ *and* $G_{K,k_1}(D,1) = F_J(D)$. □

Theorem 4.6: Let K and k_1 be as in Definition 4.8. Then the equation $G_{K,k_1}(D_{K,k_1}(\lambda),\lambda) = D_{K,k_1}(\lambda)$ defines D_{K,k_1} as a continuous, continuously differentiable function $[0,1] \to \mathbb{R}$ unless $\exists k \in K$ such that $\nu_k = 0$, when D_{K,k_1} may fail to be differentiable at points where it is zero. □

Proof 4.6: *P By a simple modification of Theorem 4.2,* $\forall\lambda \in [0,1]$, $\exists! D_\lambda$ *such that* $G_{K,k_1}(D_\lambda,\lambda) - D_\lambda = 0$, *which is enough to define the function* $D_{K,k_1}(\lambda)$ *such that* $G_{K,k_1}(D_{K,k_1}(\lambda),\lambda) - D_{K,k_1}(\lambda) = 0$.

To show that D_{K,k_1} *is differentiable, we need the further facts that* $H_{K,k_1} : (D,\lambda) \mapsto G_{K,k_1}(D,\lambda) - D$, *is continuously differentiable on* $\mathbb{R}\times\mathbb{R}$ *(*$[\mathbb{R} - \{0\}] \times \mathbb{R}$ *if* $\exists k \in K$ *such that* $\nu_k = 0$*) and,*

$$\frac{\partial H_{K,k_1}(D,\lambda)}{\partial D} = \frac{1}{p}\sum_{k\in K} D\left(D^2 + \frac{4\nu_k^2}{g_k^2}\right)^{-\frac{1}{2}} - \frac{\lambda}{p} D\left(D^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}\right)^{-\frac{1}{2}} - 1 < 0$$

on $\mathbb{R}\times[-\epsilon, 1+\epsilon] \supset \mathbb{R}\times[0,1]$ *(or* $[\mathbb{R}-\{0\}]\times[-\epsilon,1+\epsilon] \supset [\mathbb{R}-\{0\}]\times[0,1]$ *if* $\exists k \in K$ *such that* $\nu_k = 0$*), where* ϵ *is a sufficiently small positive constant (as* $(\overline{\overline{K}} - \lambda)/p \leq 1 \forall \lambda \in [0,1]$*). Hence, by the implicit function theorem (see Burkhill and Burkill [4]), the equations*

$$D_{\lambda_0}(\lambda) = G_{K,k_1}(D_{\lambda_0}(\lambda),\lambda), \quad D_{\lambda_0}(\lambda_0) = G_{K,k_1}(D_{\lambda_0}(\lambda_0),\lambda_0),$$

define D_{λ_0}, *for each* $\lambda_0 \in [0,1]$, *as a continuously differentiable function of* λ, *on some nonempty closed interval* $[\lambda_0 - \delta_{\lambda_0}, \lambda_0 + \delta_{\lambda_0}]$, *except when* $\exists k \in K$ *such that* $\nu_k = 0$, *when this only remains necessarily true so long as* $D_{\lambda_0}(\lambda_0)$ *remains nonzero. But, obviously,* $D_{K,k_1}(\lambda) = D_{\lambda_0}(\lambda)$ *on each* $[\lambda_0 - \delta_{\lambda_0}, \lambda_0 + \delta_{\lambda_0}]$, *and* D_{K,k_1} *is itself continuously differentiable, with the possible exception of*

[1] Of course, we necessarily have $\Lambda_K(A) = \Lambda(q)$ only when $A = A_K$ and $q = \hat{q}_K$, as otherwise equations (4.32) and (4.35) need not hold.

points where D_{K,k_1} is zero if $\exists k \in K$ such that $\nu_k = 0$, when D_{K,k_1} is still continuous, as a consequence of the continuity of $G_{K,k_1} - \iota$ ($\iota : \mathbb{R} \to \mathbb{R} : D \mapsto D$) and the uniqueness of the solution of $G_{K,k_1}(D,\lambda) = D \ \ \forall \lambda \in [0,1]$. ■

Definition 4.9: *Let $k_1 \in K \subset \mathbb{N}_s$, $K \neq \{k_1\}$. We implicitly define the function $D_{K,k_1} : [0,1] \to \mathbb{R}$ whose existence was proved in Theorem 4.6:*

$$D_{K,k_1}(\lambda) = \frac{1}{p}\left[1 - \sum_{k\in K}\frac{1+\nu_k^2}{g_k} + \sum_{k\in K}\left(D_{K,k_1}^2 + \frac{4\nu_k^2}{g_k^2}\right)^{1/2} - \lambda\left\{\left(D_{K,k_1}^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}\right)^{1/2} - \frac{1+\nu_{k_1}^2}{g_{k_1}}\right\}\right].$$

Then $D_{K,k_1}(0) = A_K$, $D_{K,k_1}(1) = A_{K-\{k_1\}}$ and D_{K,k_1} is continuous, and is continuously differentiable except, possibly, at points where it is zero. □

Theorem 4.7: Let $k_1 \in K \subset \mathbb{N}_s$, $K \neq \{k_1\}$. Then, if

- $D_{K,k_1}(0) \in (-\infty, -|1-\nu_{k_1}^2|/g_{k_1})$, $D_{K,k_1}(\lambda) \in (-\infty, -|1-\nu_{k_1}^2|/g_{k_1})$ and D_{K,k_1} is a decreasing function $\forall \lambda \in [0,1]$;
- if $D_{K,k_1}(0) \in (-|1-\nu_{k_1}^2|/g_{k_1}, |1-\nu_{k_1}^2|/g_{k_1})$, $D_{K,k_1}(\lambda) \in (-|1-\nu_{k_1}^2|/g_{k_1}, |1-\nu_{k_1}^2|/g_{k_1})$ and D_{K,k_1} is an increasing function $\forall \lambda \in [0,1]$;
- if $D_{K,k_1}(0) \in (|1-\nu_{k_1}^2|/g_{k_1}, \infty)$, $D_{K,k_1}(\lambda) \in (|1-\nu_{k_1}^2|/g_{k_1}, \infty)$ and D_{K,k_1} is a decreasing function $\forall \lambda \in [0,1]$;
- and if $D_{K,k_1}(0) \in \{-|1-\nu_{k_1}^2|/g_{k_1}, |1-\nu_{k_1}^2|/g_{k_1}\}$, D_{K,k_1} is constant.

□

Proof 4.7: *P By the definition of D_{K,k_1},*

$$\frac{\mathrm{d}D_{K,k_1}}{\mathrm{d}\lambda} = \frac{\partial G_{K,k_1}}{\partial D_{K,k_1}}\frac{\mathrm{d}D_{K,k_1}}{\mathrm{d}\lambda} + \frac{1}{p}\left[\frac{1+\nu_{k_1}^2}{g_{k_1}} - \sqrt{D_{K,k_1}(\lambda)^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}}\right],$$

so

$$\frac{\mathrm{d}D_{K,k_1}}{\mathrm{d}\lambda} = \frac{\frac{1+\nu_{k_1}^2}{g_{k_1}} - \sqrt{D_{K,k_1}(\lambda)^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}}}{p\left(1 - \frac{\partial G_{K,k_1}}{\partial D_{K,k_1}}\right)}, \tag{4.41}$$

where, of course,

$$\frac{\partial G_{K,k_1}}{\partial D_{K,k_1}} = \frac{1}{p}\left[\sum_{k\in K} D_{K,k_1}(\lambda)\left(D_{K,k_1}(\lambda)^2 + \frac{4\nu_k^2}{g_k^2}\right)^{-\frac{1}{2}} + D_{K,k_1}(\lambda)\left(D_{K,k_1}(\lambda)^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}\right)^{-\frac{1}{2}}\right]$$

so the denominator in equation (4.41) is in $(0, 2p)$.

Suppose $D_{K,k_1}(0) \in (-|1-\nu_{k_1}^2|/g_{k_1}, |1-\nu_{k_1}^2|/g_{k_1})$. *Then*

$$0 < \frac{\mathrm{d}D_{K,k_1}}{\mathrm{d}\lambda} < \alpha \left[\frac{1+\nu_{k_1}^2}{g_{k_1}} - \sqrt{D_{K,k_1}(\lambda)^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}} \right],$$

for some $\alpha > 0$, *so long as* $D_{K,k_1}(\lambda) \in (-|1-\nu_{k_1}^2|/g_{k_1}, |1-\nu_{k_1}^2|/g_{k_1})$. *Suppose* $\exists\lambda$ *such that* $D_{K,k_1}(\lambda) = |1-\nu_{k_1}|/g_{k_1}$. *By the continuity of* D_{K,k_1}, *we may choose* λ_0 *to be the least such* $\lambda \in [0,1]$. *Then, for* $\lambda \in [0, \lambda_0)$, D_{K,k_1} *is an increasing function, and*

$$\lambda_0 > \frac{1}{\alpha} \int_{D_{K,k_1}(0)}^{|1-\nu_{k_1}^2|/g_{k_1}} \left(\frac{1+\nu_{k_1}^2}{g_{k_1}} - \sqrt{D^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}} \right)^{-1} \mathrm{d}D.$$

But the right-hand side of this inequality is unbounded, so $\lambda_0 \notin \mathbb{R} \supset [0,1]$.

Hence $D_{K,k_1}(\lambda) \in (-|1-\nu_{k_1}^2|/g_{k_1}, |1-\nu_{k_1}^2|/g_{k_1})$ $\forall\lambda \in [0,1]$ *and the proof for the rest of the Theorem follows the same lines.* ■

The conclusion of Theorem 4.7 is illustrated in Figure (4.4).

Theorem 4.8: Let $\mathbb{N}_s \supset K = J \cup \{k_1\}$, $k_1 \notin K$, $J \neq \emptyset$ and suppose A_K is valid. Then, $\Lambda(\hat{q}_K) = \Lambda_K(A_K) < \Lambda(\hat{q}_J) = \Lambda_J(A_J)$. □

Proof 4.8: *P We define* M *by*

$$\begin{aligned} M(\lambda) &= M(\lambda, D_{K,k_1}(\lambda)) \\ &= p \ln p D_{K,k_1}(\lambda) - \sum_{k\in K} \frac{1}{2} \ln \left[g_k D_{K,k_1}(\lambda) + \sqrt{g_k^2 D_{K,k_1}(\lambda)^2 + 4\nu_k^2} \right] \\ &\quad + \lambda \ln \frac{1}{2} \left[g_{k_1} D_{K,k_1}(\lambda) + \sqrt{g_{k_1}^2 D_{K,k_1}(\lambda)^2 + 4\nu_{k_1}^2} \right]. \end{aligned} \tag{4.42}$$

Then $D_{K,k_1}(0) = A_K$, $D_{K,k_1}(1) = A_J$, $e^{M(0)} = \Lambda_K(A_K)$, $e^{M(1)} = \Lambda_J(A_J)$, *and*

$$\begin{aligned} \frac{\mathrm{d}M(\lambda)}{\mathrm{d}\lambda} &= \frac{1}{D_{K,k_1}(\lambda)} \left[p - \sum_{k\in K} \frac{D_{K,k_1}(\lambda)}{\sqrt{D_{K,k_1}(\lambda)^2 + \frac{4\nu_k^2}{g_k^2}}} + \lambda \frac{D_{K,k_1}(\lambda)}{\sqrt{D_{K,k_1}(\lambda)^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}}} \right] \frac{\mathrm{d}D_{K,k_1}(\lambda)}{\mathrm{d}\lambda} \\ &\quad + \ln[g_{k_1} D_{K,k_1}(\lambda) + \sqrt{g_{k_1}^2 D_{K,k_1}(\lambda)^2 + 4\nu_{k_1}^2}] \\ &= \frac{1}{D_{K,k_1}(\lambda)} \left[\frac{1+\nu_{k_1}^2}{g_{k_1}} - \sqrt{D_{K,k_1}(\lambda)^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}} \right] \\ &\quad + \ln[g_{k_1} D_{K,k_1}(\lambda) + \sqrt{g_{k_1}^2 D_{K,k_1}(\lambda)^2 + 4\nu_{k_1}^2}]. \end{aligned}$$

Consider $f(x) = x^{-1}[b - \sqrt{x^2 + a^2}] + \ln[x + \sqrt{x^2 + a^2}]$, $b > |a|$. *On* $(0, \infty)$, *this has a minimum at* $x = \sqrt{b^2 - a^2}$ *(by the usual process of differentiation and setting to zero). Hence, on putting* $x = D_{K,k_1}(\lambda)$, $b = (1 + \nu_{k_1}^2)/g_{k_1}$ *and* $a = 2|\nu_{k_1}|/g_{k_1}$, *we have, for nonnegative* $D_{K,k_1}(\lambda)$,

$$\begin{aligned} \frac{\mathrm{d}M(\lambda)}{\mathrm{d}\lambda} &\geq \ln\left[\frac{1+\nu_{k_1}^2}{g_{k_1}} + \frac{|1-\nu_{k_1}^2|}{g_{k_1}}\right] + \ln g_{k_1} \\ &= \ln[1 + \nu_{k_1}^2 + |1 - \nu_{k_1}^2|] > 0. \end{aligned} \tag{4.43}$$

Now $D_{K,k_1}(0) = A_K \geq (1 - \nu_{k_1}^2)/g_{k_1}$ *and* $D_{K,k_1}(0) > 0$, *so, either* $D_{K,k_1}(0) \geq |1 - \nu_{k_1}^2|/g_{k_1}$, *and* $D_{K,k_1}(\lambda) \geq |1 - \nu_{k_1}^2|/g_{k_1} \geq 0$ *by Theorem 4.7, or* $|1 - \nu_{k_1}^2|/g_{k_1} > D_{K,k_1}(0) \geq -|1 - \nu_{k_1}^2|/g_{k_1}$, *and* D_{K,k_1} *is increasing or constant, again by Theorem 4.7, on* $[0, 1]$, *so* $D_{K,k_1}(\lambda) > 0$. *If* $\nu_{k_1}^2 = 1$, *the first alternative leaves open the possibility that* $D_{K,k_1}(\lambda) = 0$ *for some* λ, *but this is not possible as* $A_K > 0$ *implies* $D_{K,k_1}(0) > |1 - \nu_{k_1}^2|/g_{k_1} = 0$, *so* $D_{K,k_1}(\lambda) > 0 \ \forall \lambda \in [0, 1]$, *again by Theorem 4.7. Consequently,* $D_{K,k_1}(\lambda)$ *is nonnegative for* $\lambda \in [0, 1]$ *and so inequality (4.43) means that* $\Lambda_J(A_J) > \Lambda_K(A_K)$. ■

As an immediate corollary we have:

Theorem 4.9: Let $\mathbb{N} \supset K \supsetneq J \neq \emptyset$ and suppose A_K is valid. Then $\Lambda(\hat{q}_K) = \Lambda_K(A_K) < \Lambda(\hat{q}_J) = \Lambda_J(A_J)$. □

Proof 4.9: *P A simple consequence of Theorem 4.8, as the elements of* $K - J$ *can be removed from* K *one by one.* ■

Our next task is to find out what happens when we start from a set K such that either A_K is undefined or is invalid.

Theorem 4.10: If $s = p$ and $\sum_{k=1}^{p} \frac{1+\nu_k^2}{g_k} \leq 1$ (so $A_{\mathbb{N}_p}$ is undefined), and $K \subsetneq \mathbb{N}_p$, then $\Lambda_K(A_K) > \Lambda_\infty$. □

Proof 4.10: *P Suppose first that* $\nu_k = 0 \forall k \in \mathbb{N}_p$. *Then*

$$\Lambda_K(A_K) = \frac{p^p A_K^p}{\prod_{k \in K} g_k A_K} = \frac{p^p A_K^{p - \overline{\overline{K}}}}{\prod_{k \in K} g_k}.$$

But

$$A_K = F_K(A_K) = \frac{1}{p}\left[1 - \sum_{k \in K} \frac{1}{g_k} + \overline{\overline{K}} A_K\right]$$

which implies that $A_K = 1 - \frac{1}{p - \overline{\overline{K}}} \sum_{k \in K} g_k^{-1}$ *and*

$$\Lambda_K(A_K) = \frac{p^p}{\prod_{k \in K} g_k} \left(\frac{1 - \sum_{k \in K} g_k^{-1}}{p - \overline{\overline{K}}}\right)^{p - \overline{\overline{K}}}$$

$$\geq \frac{p^p}{\prod_{k\in K} g_k}\left(\frac{\sum_{k\in\mathbb{N}_p-K} g_k^{-1}}{p-\overline{\overline{K}}}\right)^{p-\overline{\overline{K}}}$$
$$\geq \frac{p^p}{\prod_{k=1}^{p} g_k} = \Lambda_\infty,$$

as the arithmetic mean of a set of positive numbers always equals or exceeds their geometric mean, and we also have $1-\sum_{k=1}^{p} g_k^{-1} \geq 0$ *implying* $1-\sum_{k\in K} g_k^{-1} \geq \sum_{k\in\mathbb{N}_p-K} g_k^{-1}$.

Now suppose $\exists k_2 \in \mathbb{N}_p$ *such that* $\nu_{k_2} \neq 0$.

Let

$$G(D,\lambda) = \frac{1}{p}\left[1-\sum_{k=0}^{p}\frac{1+\nu_k^2}{g_k}+\sum_{k=0}^{p}\left(D^2+\frac{4\nu_k^2}{g_k^2}\right)^{1/2} -\lambda\left\{\sum_{k\in K}\left(D^2+\frac{4\nu_k^2}{g_k^2}\right)^{1/2}-\sum_{k\in K}\frac{1+\nu_k^2}{g_k}\right\}\right].$$

Then, by the arguments of Theorem 4.8, the equation $D(\lambda)=G(D(\lambda),\lambda)$ *defines* D *as a continuously differentiable positive function on* $[\lambda_0,1]\forall\lambda_0\in(0,1)$ *and so* D *is defined as a continuously differentiable positive function on* $(0,1]$.

Now we prove that $\lim_{\lambda\to 0+} D(\lambda)=\infty$.

We have that

$$D(\lambda) > \frac{p-\overline{K}\lambda}{p}D(\lambda)+\frac{\lambda}{p}\sum_{k\in K}\frac{1+\nu_k^2}{g_k},$$

as $\left(D(\lambda)^2+\frac{4\nu_{k_2}^2}{g_{k_2}}\right)^{1/2} > D(\lambda)$, *so*

$$D(\lambda) > \frac{1}{\overline{\overline{K}}}\sum_{k\in K}\frac{1+\nu_k^2}{g_k} \quad \forall\lambda\in(0,1].$$

Since $D(\lambda)$ *is bounded away from zero and* $\nu_{k_2}\neq 0$, $\exists\alpha>0$ *such that* $\sqrt{D(\lambda)^2+\frac{4\nu_{k_2}^2}{g_{k_2}}} > D(\lambda)+\frac{\alpha}{D(\lambda)}$.

Then $D(\lambda) > \frac{p-\overline{\overline{K}}\lambda}{p}D(\lambda)+\frac{\lambda}{p}\sum_{k\in K}\frac{1+\nu_k^2}{g_k}+\frac{\alpha}{pD(\lambda)}$, *so*

$$D(\lambda) > \frac{1}{\overline{\overline{K}}}\sum_{k\in K}\frac{1+\nu_k^2}{g_k}+\frac{\alpha}{\overline{\overline{K}}\lambda D(\lambda)}$$

and if $\lim_{\lambda\to 0+}\lambda D(\lambda)=0$ *holds, the right-hand side of this inequality tends to* ∞, *so* $D(\lambda)\to\infty$; *and if* $\lim_{\lambda\to 0+}\lambda D(\lambda)=0$ *does not hold, we may conclude that* $D(\lambda)\to\infty$ *if we may assume that* D *is monotonic.*

But

$$\frac{\mathrm{d}D}{\mathrm{d}\lambda}=\frac{\sum_{k\in K}\frac{1+\nu_k^2}{g_k}-\sum_{k\in K}\sqrt{D(\lambda)^2+\frac{4\nu_k^2}{g_k^2}}}{p\left(1-\frac{\partial g_k}{\partial D}\right)}<0$$

on $(0,1]$, *as* $D(\lambda) > \frac{1}{\overline{\overline{K}}} \sum_{k \in K} \frac{1+\nu_k^2}{g_k}$ *on* $(0,1]$ *and* $\frac{\partial G}{\partial D} < 1$. *Hence,* $\lim_{\lambda \to 0+} D(\lambda) = \infty$.

Also,

$$D(\lambda) < \frac{1}{p}\left[1 - \sum_{k=1}^{p} \frac{(1-|\nu_k|)^2}{g_k} + pD(\lambda) - \lambda\left\{\overline{\overline{K}} D(\lambda) - \sum_{k \in K} \frac{1+\nu_k^2}{g_k}\right\}\right],$$

by the definition of D, *so*

$$D(\lambda) < \frac{1}{\overline{\overline{K}}\lambda}\left(1 - \sum_{k=1}^{p} \frac{(1-|\nu_k|)^2}{g_k}\right) + \frac{1}{\overline{\overline{K}}} \sum_{k \in K} \frac{1+\nu_k^2}{g_k}. \tag{4.44}$$

Defining M *by*

$$\begin{aligned} M(\lambda) &= p \ln pD(\lambda) - \sum_{k=1}^{p} \frac{1}{2} \ln \left[g_k D(\lambda) + \sqrt{g_k^2 D(\lambda)^2 + 4\nu_k^2}\right] \\ &\quad + \lambda \sum_{k \in K} \ln \frac{1}{2}\left[g_k D(\lambda) + \sqrt{g_k^2 D(\lambda)^2 + 4\nu_k^2}\right], \end{aligned}$$

we have that

$$\begin{aligned} M(\lambda) &= p \ln p - \sum_{k=1}^{p} \frac{1}{2} \ln \left[g_k + \sqrt{g_k^2 + \frac{4\nu_k^2}{D(\lambda)^2}}\right] \\ &\quad + \lambda \sum_{k \in K} \ln \frac{1}{2}\left[g_k D(\lambda) + \sqrt{g_k^2 D(\lambda)^2 + 4\nu_k^2}\right], \end{aligned}$$

so

$$\begin{aligned} \lim_{\lambda \to 0+} M(\lambda) &= p \ln p - \sum_{k=1}^{p} \ln g_k \\ &\quad + \lim_{\lambda \to 0+} \lambda \sum_{k \in K} \ln \frac{1}{2}\left[g_k D(\lambda) + \sqrt{g_k^2 D(\lambda)^2 + 4\nu_k^2}\right] \\ &= p \ln p - \sum_{k=1}^{p} \ln g_k, \end{aligned}$$

by equation (4.44). Thus $\lim_{\lambda \to 0+} e^{M(\lambda)} = \Lambda_\infty$, *and, of course,* $e^{M(1)} = \Lambda_K(A_K)$.

By the argument of Theorem 4.8, $\frac{\mathrm{d}M(\lambda)}{\mathrm{d}\lambda} > 0$ *as we know that* $D(\lambda) > \frac{1}{\overline{\overline{K}}} \sum_{k \in K} \frac{1+\nu_k^2}{g_k} > 0$, *and so* $\Lambda_\infty = \lim_{\lambda \to 0+} e^{M(\lambda)} < e^{M(1)} = \Lambda_K(A_K)$. ■

We now wish to establish some facts which will enable us to find the set $K \subset \mathbb{N}_s$ such that A_K is valid and $\Lambda_K(A_K) <= \Lambda_J(A_J)$ for all $J \subset \mathbb{N}_s$ such that A_J is valid.

Theorem 4.11: If $K \subset \mathbb{N}_s$, $k_1, k_2 \in \mathbb{N}_s - K$, $k_1 \neq k_2$, $\frac{1-\nu_{k_2}^2}{g_{k_2}} \geq \max_{k \in K \cup \{k_1\}} \frac{1-\nu_k^2}{g_k}$, $|\nu_{k_1}| \leq 1$ and $A_{K \cup \{k_2\}}$ is valid, then $A_{K \cup \{k_1\}}$ is valid. □

Proof 4.11: *P Assume that* $A_{K \cup \{k_1\}}$ *is not valid. Then*

$$A_{K \cup \{k_1\}} < \max_{k \in K \cup \{k_1\}} \frac{1-\nu_k^2}{g_k} \leq \frac{1-\nu_{k_2}^2}{g_{k_2}} \leq A_{K \cup \{k_2\}},$$

so, by Theorem 4.3, $F_{K\cup\{k_2\}}\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right) \geq \frac{1-\nu_{k_2}^2}{g_{k_2}} > F_{K\cup\{k_1\}}\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right)$ *and consequently,*

$$\begin{aligned} p\left(F_{K\cup\{k_2\}}\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right)\right. & \\ \left. - F_{K\cup\{k_1\}}\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right)\right) &= p\left(F_K\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right) - F_{K\cup\{k_1\}}\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right)\right) \\ &= \frac{1+\nu_{k_1}^2}{g_{k_1}} - \sqrt{\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right)^2 + \frac{4\nu_{k_1}^2}{g_{k_1}^2}} > 0. \end{aligned} \tag{4.45}$$

But $|\nu_{k_1}| \leq 1$ *implies that* $\frac{1-\nu_{k_2}^2}{g_{k_2}} \geq \frac{1-\nu_{k_1}^2}{g_{k_1}} \geq 0$, *and* $\left(\frac{1-\nu_{k_2}^2}{g_{k_2}}\right)^2 \geq \left(\frac{1-\nu_{k_1}^2}{g_{k_1}}\right)^2$. *But this means that equation (4.45) cannot hold, a contradiction. Thus, the Theorem must hold.* ■

From this point on, we will assume that our hyperplanes are labelled so that

$$\frac{1-\nu_1^2}{g_1} \leq \frac{1-\nu_2^2}{g_2} \leq \cdots \leq \frac{1-\nu_s^2}{g_s}. \tag{4.46}$$

Conjecture 4.12: If $A_{\mathbb{N}_r}$ is valid and $A_{\mathbb{N}_k}$ is invalid $\forall k > r$, then $\Lambda_{\mathbb{N}_r}(A_{\mathbb{N}_r}) \leq \Lambda_K(A_K) \forall K \subset \mathbb{N}_s$.

□

Conjecture 4.12 enables the production of Algorithm 4.1 finding an ellipsoid containing the

intersection of an ellipsoid $\mathcal{E}(a,Q)$ and s Q-orthogonal pairs of parallel hyperplanes.

Algorithm: 4.1

1. Given $\mathcal{E}(a,Q)$ and hyperplane pairs $\mathbb{H}_1^{\pm}(n_1,y_1)\ldots\mathbb{H}_s^{\pm}(n_s,y_s)$,

2. Calculate $g_k = {n_k}^{\mathrm{T}}Qn_k$, $\nu_k = y_k - {n_k}^{\mathrm{T}}a$, $k=1,\ldots,s$.

3. If $s=p$ and $\sum_{k=1}^{p}\frac{1+\nu_k^2}{g_k}\leq 1$, set $Q_1 = pQ(\sum_{k=1}^{p}g_k^{-2}n_k{n_k}^{\mathrm{T}})Q$ and $a_1 = Q\sum_{k=1}^{p}g_k^{-1}y_kn_k$.

   Stop.

4. (a) Redefine k so that $\frac{1-\nu_1^2}{g_1}\leq\ldots\leq\frac{1-\nu_1^2}{g_1}$;

   (b) Set $\ell = s$;

   (c) While $\ell>0$ and $\sum_{k=1}^{\ell}\frac{(1-|\nu_k|)^2}{g_\ell}>1$

      i. Calculate $A_{\mathbb{N}_\ell}$ such that $A_{\mathbb{N}_\ell} = F_{\mathbb{N}_\ell}(A_{\mathbb{N}_\ell})$;

      ii. if $A_{\mathbb{N}_\ell}\geq\frac{1-\nu_\ell^2}{g_\ell}$, break from loop and go to 4d;

      iii. else set $\ell := \ell-1$;

   (d) if $\ell>0$, set

$$
\begin{aligned}
q_i &= \frac{1}{2g_i}\left[A_{\mathbb{N}_\ell}g_i - 2 + \sqrt{A_{\mathbb{N}_\ell}^2g_i+4\nu_i^2}\right], i=1,\ldots,\ell\\
Q_1 &= \left[1+\sum_{k=1}^{\ell}\left(q_k-\frac{\nu_k^2q_k}{1+g_kq_k}\right)\right]Q\left[\mathrm{I}-\sum_{k=1}^{\ell}\frac{q_k}{1+g_kq_k}n_k{n_k}^{\mathrm{T}}\right]Q\\
a_1 &= a+Q\sum_{k=1}^{\ell}\frac{nu_kq_k}{1+g_kq_k}n_k.
\end{aligned}
$$

To make use of this, a method is needed to obtain the Q-orthogonal hyperplanes from the available hyperplane pairs.

4.3.1 Hyperplane Shifting — s Pairs at a Time

To obtain Q-orthogonal hyperplane pairs from the available hyperplanes, first shift these latter hyperplanes according to the method of subsection 4.2.1, that is, take $\mathbb{H}_1^{\pm}$, find the $\mathbb{H}_k^{\pm}$, $k\in\{2,\ldots,s\}$ such that the members of $\mathbb{H}_1^{\pm}$ can be shifted by the greatest extent to touch $\mathcal{E}\cap\Pi_k$, (where $\mathcal{E}$ is the relevant ellipsoid), redefine $\mathbb{H}_1^{\pm} = \mathbb{H}_1^{\pm}(n_1,y_1)$ to be the shifted hyperplane pair and then find $\mathbb{H}_k^{\pm}$, $k\in\{1,3,4,\ldots,s\}$ such that the members of $\mathbb{H}_2^{\pm}$ can be shifted by the

greatest extent to touch $\mathcal{E} \cap \Pi_k$, and work through all the $\mathbb{H}^{\pm}$'s in this fashion. The next step is to perform a Gram-Schmidt style orthogonalisation on the normals $n_s, n_{s-1}, \ldots, n_1$ to the hyperplanes, to obtain the Q-orthogonal set $\bar{n}_s \propto n_s, \bar{n}_{s-1}, \ldots, \bar{n}_1$:

$$\begin{aligned} \tilde{n}_{s-j} &= n_{s-j} - \sum_{k=0}^{j-1} (\bar{n}_{s-k}{}^{\mathrm{T}} Q n_{s-j}) \bar{n}_{s-k} \\ \bar{n}_{s-j} &= \frac{\tilde{n}_{s-j}}{\sqrt{\tilde{n}_{s-j}{}^{\mathrm{T}} Q \tilde{n}_{s-j}}} \end{aligned}$$

$\mathbb{H}_s^{\pm}(n_s, y_s)$ will be the first hyperplane pair of the set of Q-orthogonal hyperplane pairs. To find the remaining members of the set, find $x_{i,j}^{+} = \max\{\bar{n}_i{}^{\mathrm{T}} x : x \in \left(\mathbb{H}_j^{+}(n_j, y_j) \cup \mathbb{H}_j^{-}(n_j, y_j)\right) \cap \mathcal{E}\}$, where the methods of Subsection 4.2.1 are used to determine $x_{i,j}^{+}$, and set $x_i^{+} \in \mathcal{E} \bigcap \left(\bigcup_{j=1}^{s} \mathbb{H}_j^{+}(n_j, y_j) \cup \mathbb{H}_j^{-}(n_j, y_j)\right)$ such that $\bar{n}_i{}^{\mathrm{T}} x_i^{+} = \min_{1 \le j \le s} \bar{n}_i{}^{\mathrm{T}} x_{i,j}^{+}$.

The quantity x_i^{-} satisfies the similar conditions to x_i^{+}, but with "max" and "min" interchanged. Then the ith member of the set of Q-orthogonal hyperplane pairs $\overline{\mathbb{H}}_i^{\pm}$ will have a normal parallel to $\tilde{n}_i$ and $\overline{\mathbb{H}}_i^{+}$ will pass through x_i^{+}, $\overline{\mathbb{H}}_i^{-}$ through x_i^{-}.

These hyperplane pairs can now be used in Algorithm 4.1.

Given an initial value for the ellipsoid $\mathcal{E}(a, Q)$, the following produces a sequence of ellipsoids with decreasing volume:

Algorithm: 4.2

1. Add hyperplane pair k to the set $\{\mathbb{H}_{i_1}, \ldots, \mathbb{H}_{i_{s-1}}\}$ to obtain $\{\mathbb{H}_{i_1}, \ldots, \mathbb{H}_{i_s}\}$;
2. shift $\{\mathbb{H}_{i_1}, \ldots, \mathbb{H}_{i_s}\}$, dropping any redundant members;
3. if $\{\mathbb{H}_{i_1}, \ldots, \mathbb{H}_{i_s}\}$ has more than p members, drop the earliest member;
4. produce the set of working Q-orthogonal hyperplanes $\{\mathbb{H}'_{i_1}, \ldots, \mathbb{H}'_{i_s}\}$;
5. perform Algorithm 4.1 on $\mathcal{E}(a, Q)$ and $\{\mathbb{H}'_{i_1}, \ldots, \mathbb{H}'_{i_s}\}$;
6. if more data, go to 1;
7. stop.

		Dimension							
		2		3		4		5	
	σ noise	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$
minimum	FH	2.97	2.62	2.48	2.19	2.11	1.95	1.90	1.81
	FHss	2.34	2.28	2.45	2.17	2.08	1.90	1.86	1.76
mean	FH	5.80	2.85	4.08	2.47	3.01	2.26	2.40	1.97
	FHss	4.33	2.58	3.67	2.42	2.87	2.18	2.31	1.92
maximum	FH	101.71	3.08	13.15	2.78	5.94	2.56	4.46	2.14
	FHss	61.33	2.80	9.14	2.70	5.53	2.44	4.14	2.05

Table 4.1: Improvement in characteristic length-ratio for two hyperplane modified Fogel-Huang algorithm (strict sequence)

4.4 Empirical Results for the Modified Algorithms

4.4.1 Two Hyperplane Pair Algorithms

Figures 4.5 to 4.15 are much the same as the corresponding Figures in Chapter 3, with the addition of lines showing the minimum, mean and maximum characteristic length-ratios from the unmodified Fogel-Huang algorithm applied to the same families of data sets. (To save space, the Figures relating to the family of data sets affected by noise with the truncated normal distribution with $\sigma_t = 1/2\sqrt{3}$ have been omitted, as their features fall between those for the uniform distribution and truncated normal distribution with $\sigma_t = 1/4\sqrt{3}$. Also omitted are Figures relating to the ellipsoid centre-true parameter distance, as any improvement here is obscured by the wild variations in this quantity.)

Hyperplanes Pairs in Strict Sequence

In every case, the improvement shown by this variant of the modified Fogel-Huang algorithm over the unmodified version after 12 steps in the mean characteristic length is dominated by the improvement for the worst case, as is shown in Table 4.1.

Table 4.1 also brings out the fact, not evident in the Figures because of the scale (dictated by the need to show the effect of the first few steps of the algorithm) that the mean performance in terms of characteristic length-ratios actually improves by about 2-25% (i.e., the uncertainty in the individual parameters decreases by this amount), and, in terms of volumes, this is roughly

		Dimension							
		2		3		4		5	
	σ noise	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$
minimum	FH	2.97	2.62	2.48	2.19	2.11	1.95	1.90	1.81
	FHoe	2.18	2.10	2.38	2.16	2.06	1.90	1.84	1.75
mean	FH	5.80	2.85	4.08	2.47	3.01	2.26	2.40	1.97
	FHoe	3.96	2.45	3.50	2.37	2.82	2.14	2.27	1.88
maximum	FH	101.71	3.08	13.15	2.78	5.94	2.56	4.46	2.14
	FHoe	70.68	2.67	11.83	2.58	5.29	2.35	3.96	2.01

Table 4.2: Improvement in characteristic length-ratio for two hyperplane modified Fogel-Huang algorithm (odd/even sequence)

6-44% ("roughly", because the estimate here involves approximating mean volumes by raising the mean characteristic length to the pth power).

In addition, the improvement in the worst cases is greater than this.

Hyperplanes Pairs in Odd/Even Sequence

As expected, using hyperplane pairs more widely separated in the sequence results in an improvement on average, as illustrated in Figures 4.8 to 4.10 and confirmed in Table 4.2, although, for the worst cases in both two and three dimensions for the uniform noise distribution, this improvement is contradicted. In percentage terms, the improvement over the unmodified Fogel-Huang algorithm varies from 4-32% for the characteristic length-ratio, and 11-53% for the volume.

"Best Second" Hyperplane Pair

Here it is assumed that the available data accumulates and the "best" hyperplane pair is selected only from those previously encountered.

Nevertheless, the departure from "onlineness" enables further improvements in performance to be made. Now the improvement over the unmodified Fogel-Huang algorithm is 11-52% for the characteristic length, and approximately 44-77%.

		Dimension							
		2		3		4		5	
	σ noise	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$
minimum	FH	2.97	2.62	2.48	2.19	2.11	1.95	1.90	1.81
	FHbs	1.89	1.86	2.05	1.76	1.92	1.71	1.70	1.63
mean	FH	5.80	2.85	4.08	2.47	3.01	2.26	2.40	1.97
	FHbs	2.77	1.98	3.05	1.93	2.56	1.81	2.14	1.69
maximum	FH	101.71	3.08	13.15	2.78	5.94	2.56	4.46	2.14
	FHbs	33.68	2.22	10.88	2.15	4.09	1.99	3.39	1.77

Table 4.3: Improvement in characteristic length-ratio for two hyperplane modified Fogel-Huang algorithm (best second)

4.4.2 s-Hyperplane Pair Algorithm

As the s-hyperplane pair algorithm makes approximations (replace hyperplane pairs by Q-orthogonal ones) not made by the two-hyperplane algorithm, it seems obvious that it will produce worse results when $s = 2$, so it has not been applied to the data sets in two-dimensional parameter space.

The results for $s > 2$ are presented in Figures 4.14 and 4.15, and in Table 4.4. The average performance here seems to be slightly better than that of the two-hyperplane algorithm with strictly following the sequence of hyperplanes, and slightly worse than that of the two-hyperplane algorithm following the "odd/even" sequence, although the performance on the worst cases is much more mixed in that the maximum characteristic length of the s-hyperplane algorithm is greater than that of the strict sequence algorithm in three cases out of six, and the same is true with regard to the "odd/even" sequence (although equality to the given accuracy also holds for one case here).

4.4.3 Conclusions

Table 4.5 presents the computational expense of the algorithms considered in this section. Comparison with Table 3.7 reveals that all of these algorithms are still cheap compared with the calculation of the minimum-volume ellipsoid. The three two-hyperplane algorithms appear to have complexity which is fairly constant as the dimension of parameter space increases, but this is certainly the result of a large constant term in that complexity, which hides an increase

		Dimension					
		3		4		5	
	σ noise	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$	$1/\sqrt{3}$	$1/4\sqrt{3}$
minimum	FH	2.48	2.19	2.11	1.95	1.90	1.81
	FHsH	2.38	2.17	2.09	1.93	1.85	1.79
mean	FH	4.08	2.47	3.01	2.26	2.40	1.97
	FHsH	3.63	2.37	2.85	2.12	2.29	1.90
maximum	FH	13.15	2.78	5.94	2.56	4.46	2.14
	FHsH	11.38	2.58	5.69	2.30	4.41	2.04

Table 4.4: Improvement in characteristic length-ratio for s hyperplane modified Fogel-Huang algorithm

	Dimension							
	2		3		4		5	
Algorithm	mean	sd	mean	sd	mean	sd	mean	sd
Strict seq.	170	26	220	89	220	85	220	90
Odd/even seq	150	28	220	11	220	9.2	230	7.8
Best second	170	30	240	93	250	74	260	78
s hyperplane pair	—	—	69	2.3	130	3.5	220	6.0

Table 4.5: Mean and standard deviation of the number of kiloflops needed to calculate 12 steps of three variants of the two hyperplane pair modified Fogel-Huang algorithm and 12 steps of the s hyperplane pair modified Fogel-Huang algorithm.

with dimension.

The interesting feature is that the s-hyperplane pair algorithm is relatively cheap compared to the two-hyperplane algorithms. This is quite possibly due to the low-dimensionality of the examples allowing the orthogonalisation involved in the s-hyperplane pair algorithm to be outweighed by the fact that the solution of the equation $F_K(A_K) = A_K$ is not needed at every step, and that a comparable operation in the two-hyperplane version is needed, i.e., the solution of the 5th-degree polynomial.

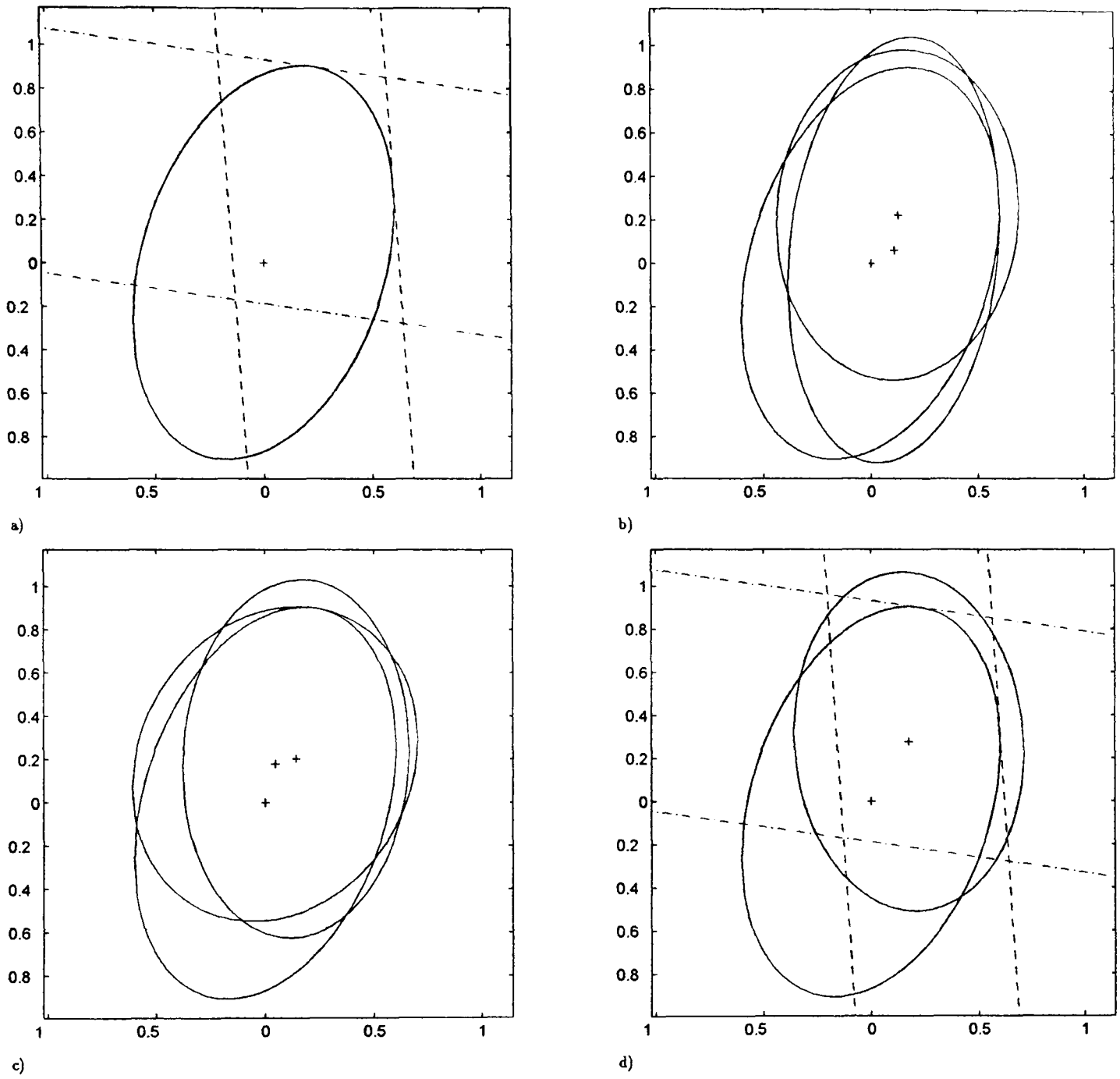


Figure 4.1: a) Initial ellipse with hyperplane pairs 1 (dashed lines) and 2 (dashed and dotted lines);
b) initial ellipse, ellipse obtained by applying the Fogel-Huang algorithm to the initial ellipse and hyperplane pair 1, and the ellipse obtained by applying Fogel-Huang to this ellipse and hyperplane pair 2 (in order of decreasing area, naturally);
c) initial ellipse, ellipse obtained by applying the Fogel-Huang algorithm to the initial ellipse and hyperplane pair 2, and the ellipse obtained by applying Fogel-Huang to this ellipse and hyperplane pair 1;
d) ellipse obtained by applying the modified algorithm to the initial ellipse and both hyperplane pairs.

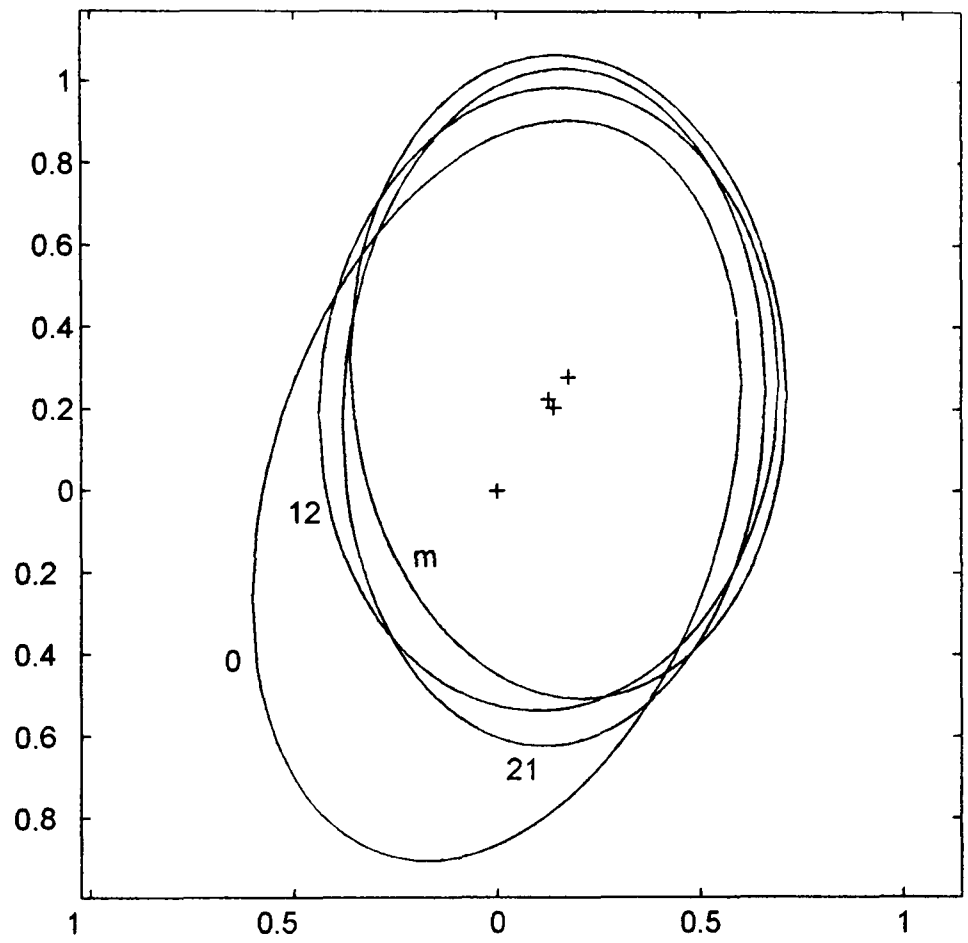


Figure 4.2: Initial ellipse ('0') with the two ellipses obtained by applying Fogel-Huang twice with the hyperplane pairs in the different orders ('12' and '21'), and the ellipse ('m') obtained by applying the modified Fogel-Huang algorithm to the initial ellipse and the two hyperplane pairs simultaneously (in this instance, the ellipse from the modified algorithm is of the smallest area).

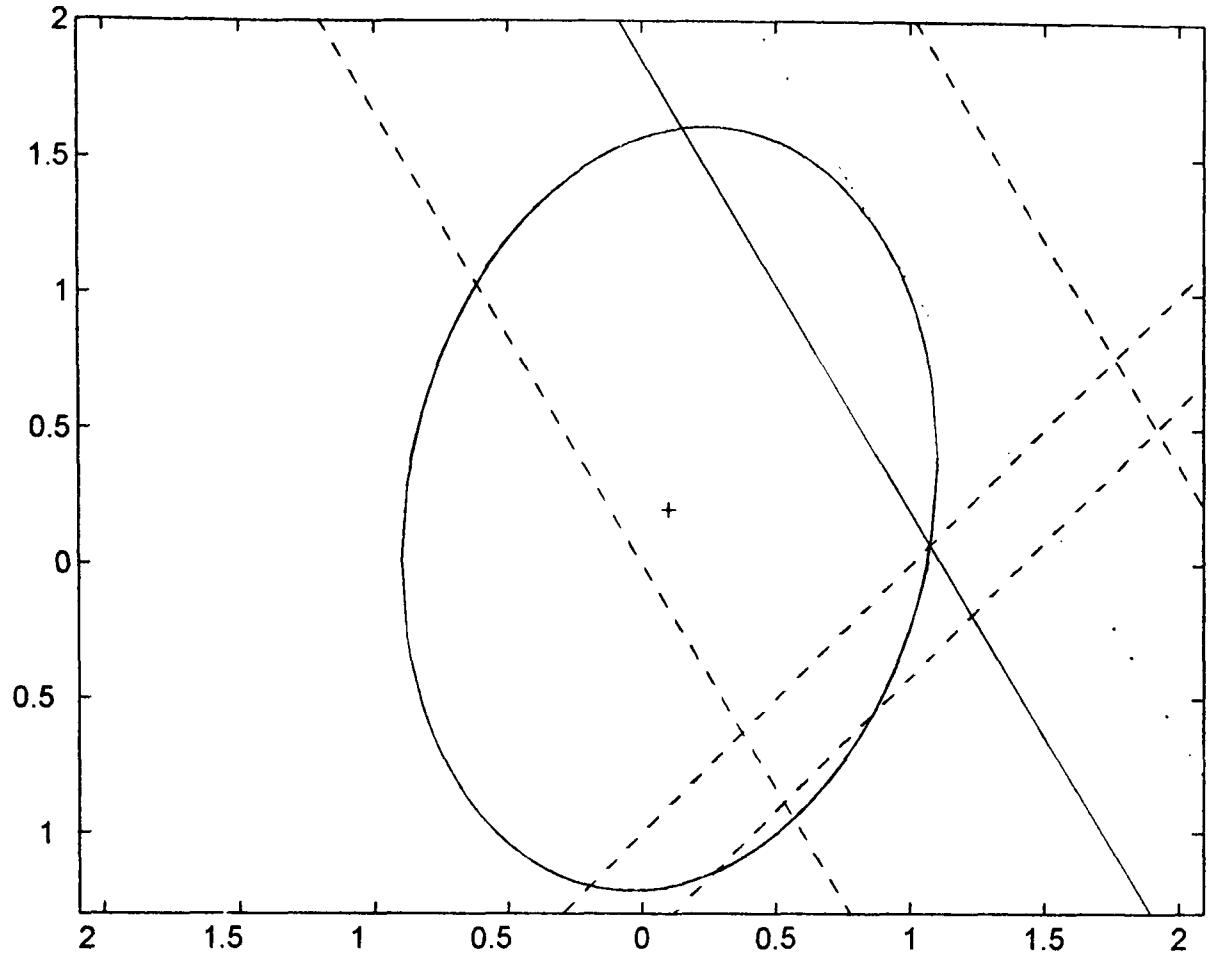


Figure 4.3: Shifting $\mathbb{H}_2^+$ to the boundary of $\mathcal{E} \cap \Pi_1$ instead of the boundary of $\mathcal{E}$ (dashed lines are the original positions of $\mathbb{H}_{1,2}^{\pm}$, the dotted line is the position of $\mathbb{H}_2^+$ after being shifted to be tangent to $\mathcal{E}$, and the solid line is the position of $\mathbb{H}_2^+$ after being shifted to touch $\Pi_1 \cap \mathcal{E}$).

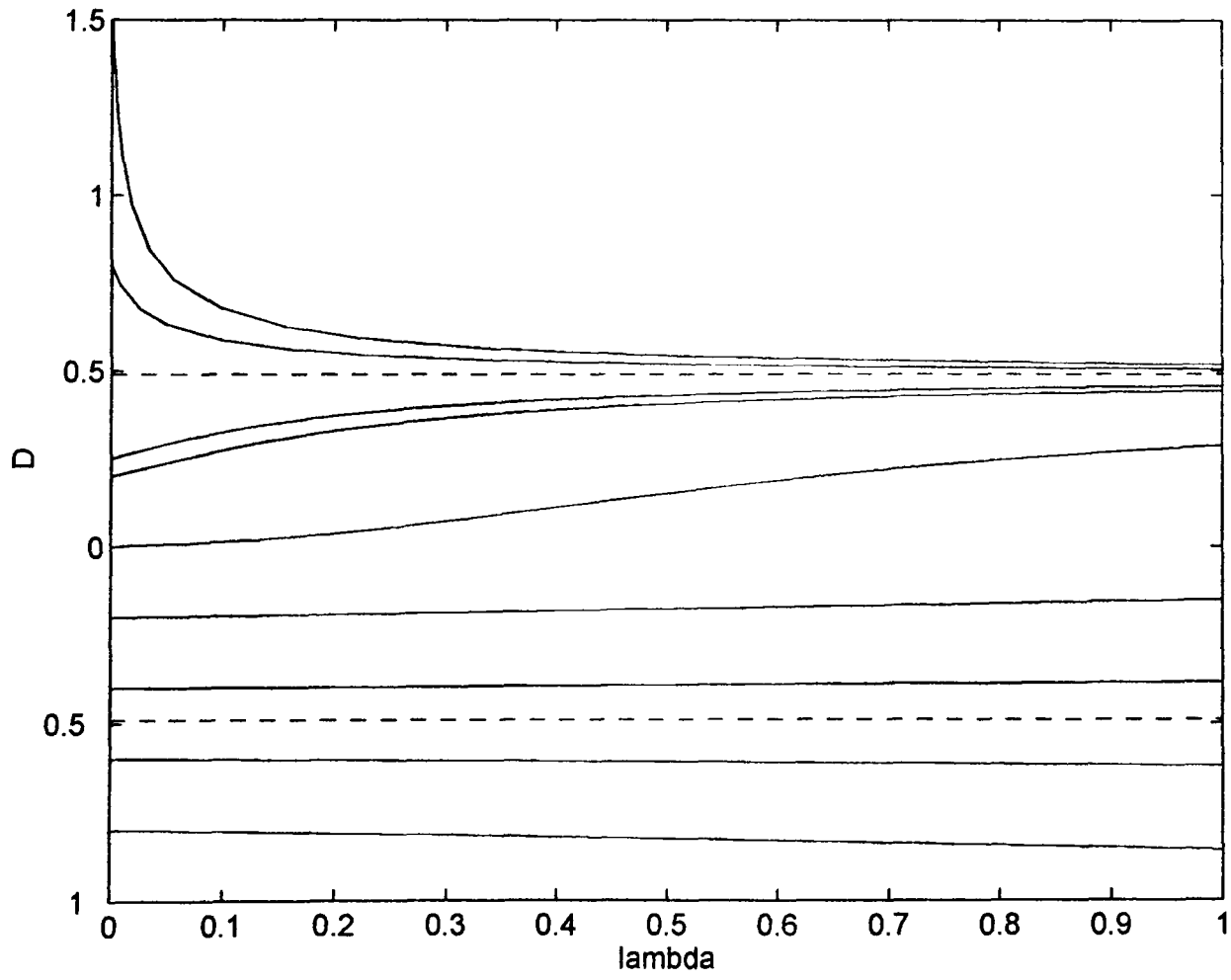


Figure 4.4: $D_{K,k_1}(\lambda)$ as the solution of Equation (4.41) of Theorem 4.7. (The dashed lines are at $D_{K,k_1} = \pm|1-\nu_{k_1}^2|/g_{k_1}$, $p=3$, $k_1=1$, $(g_1,g_2,g_3)=(1.96,2,2)$, and $(\nu_1,\nu_2,\nu_3)=(0.2,0.02,0.02)$.)

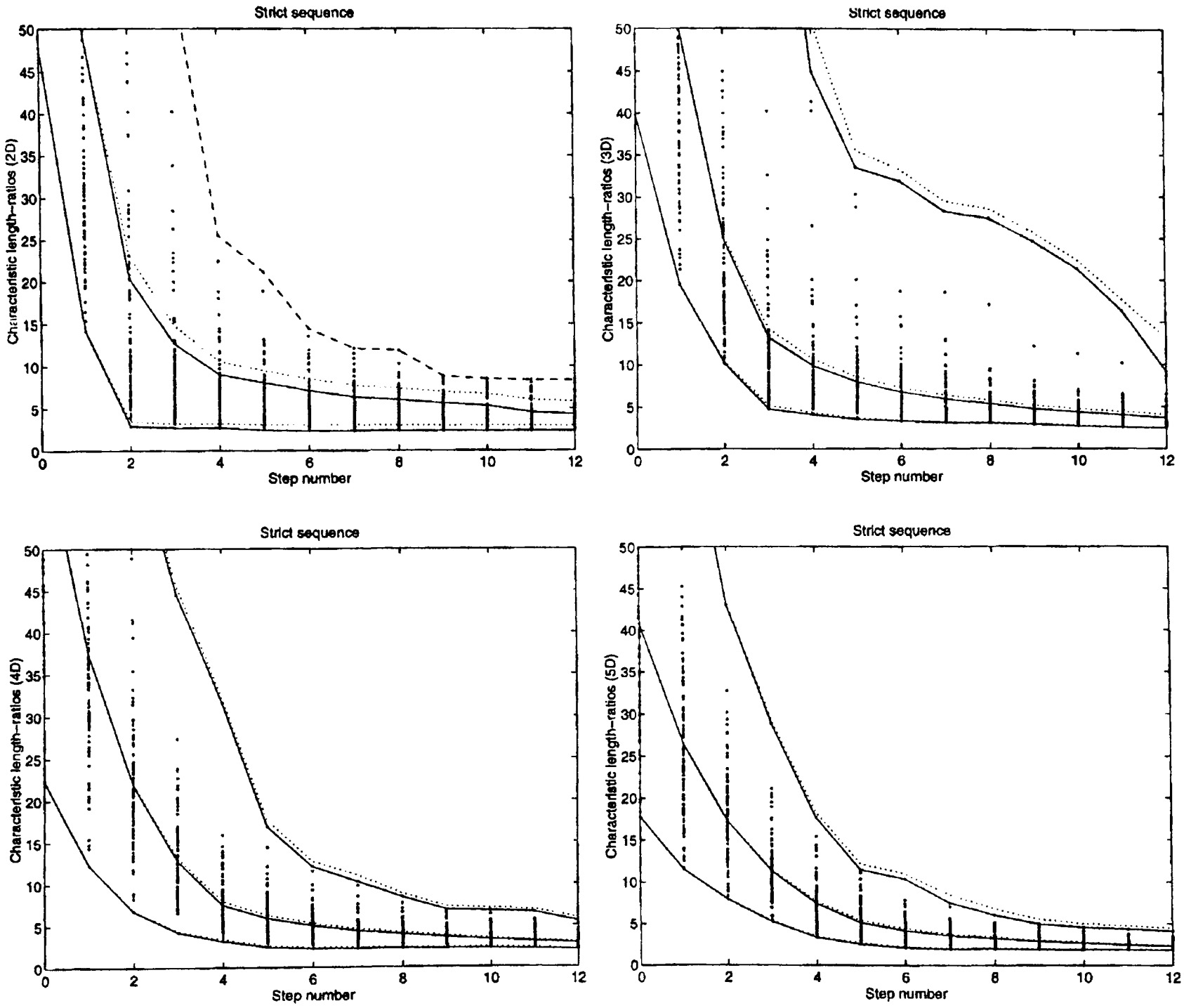


Figure 4.5: Characteristic length-ratios of strict sequence ellipsoids (noise uniformly distributed).

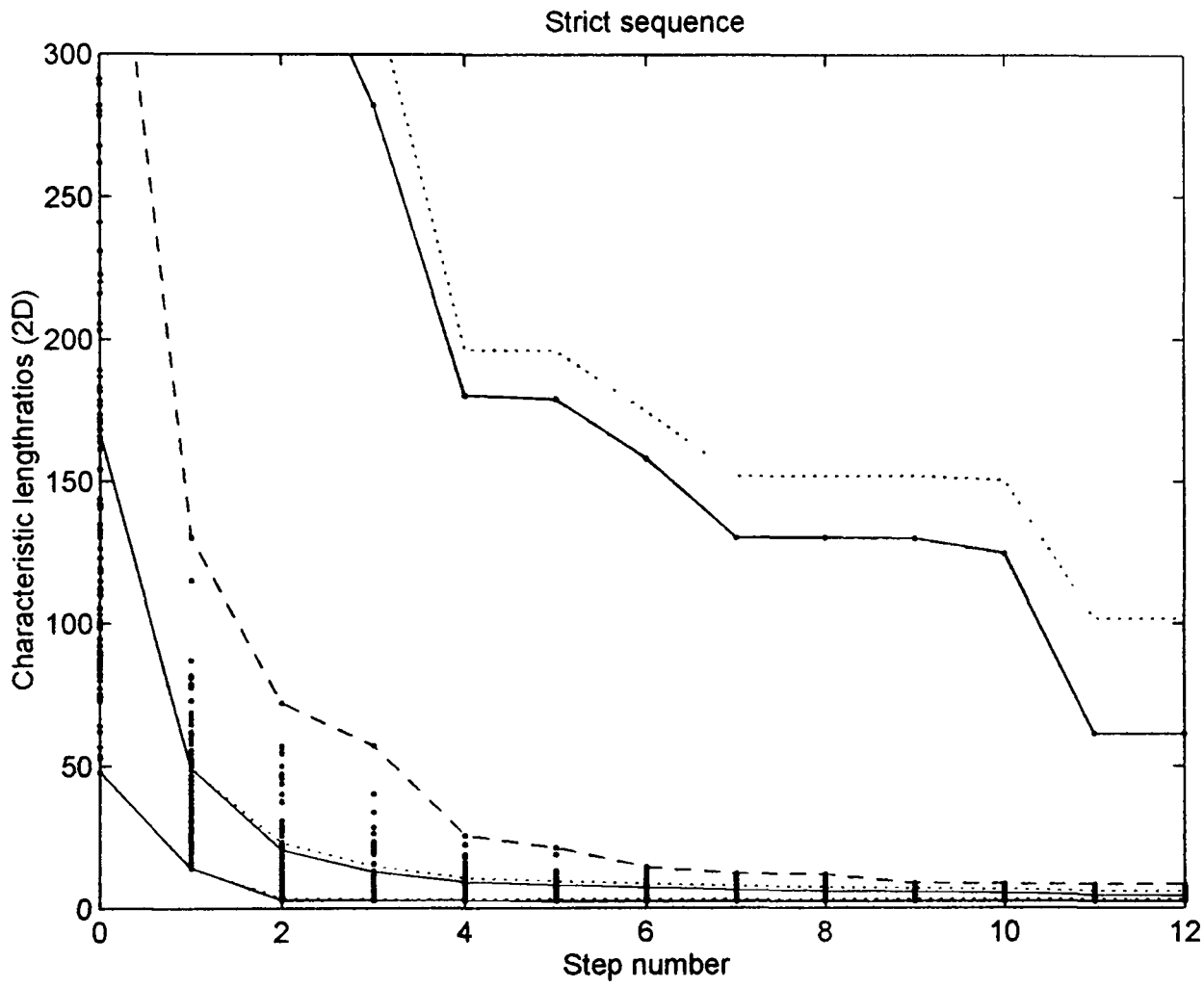


Figure 4.6: Characteristic length-ratios of strict sequence ellipsoids (noise uniformly distributed). (Showing the improvement for the worst case.)

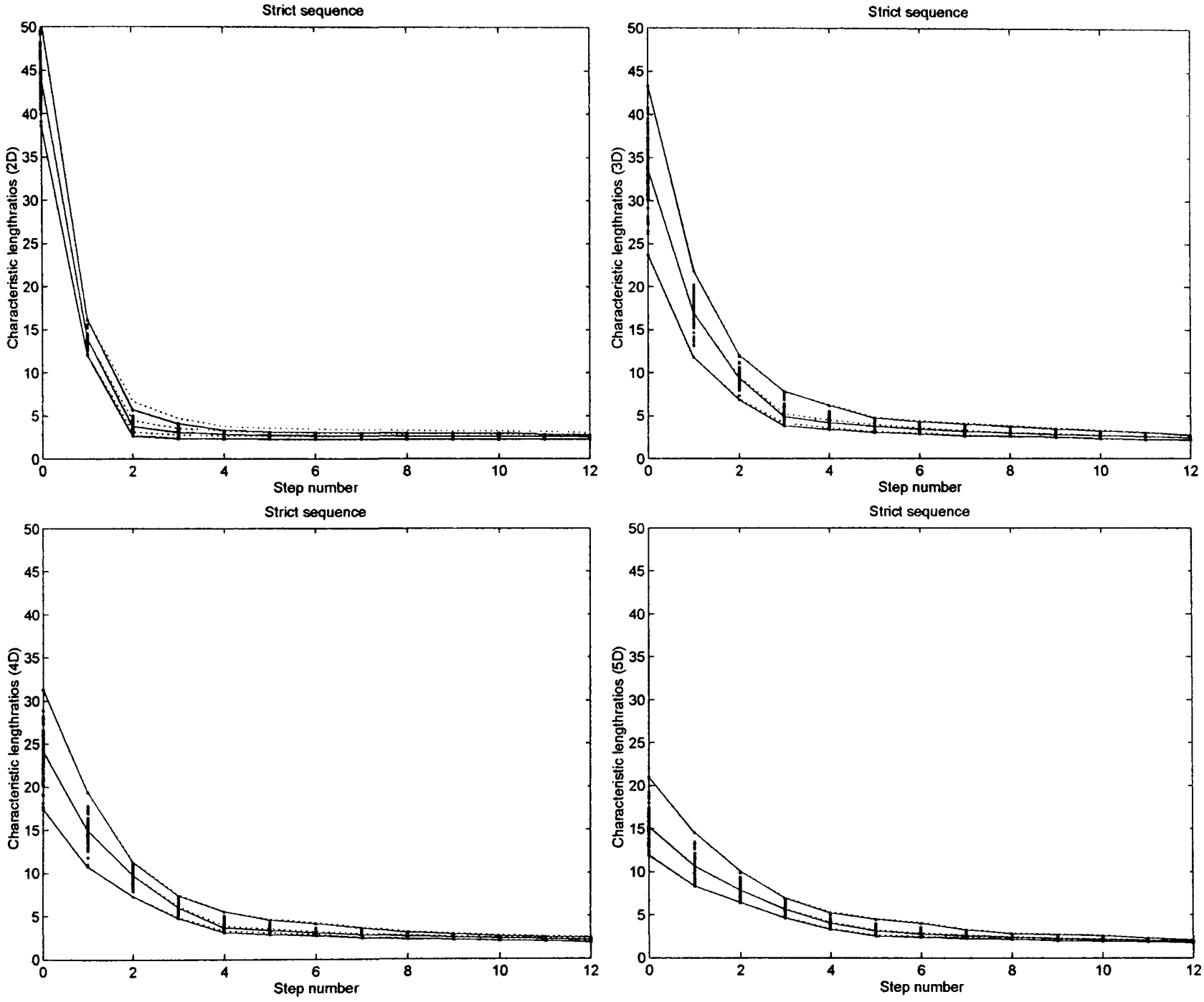


Figure 4.7: Characteristic length-ratios of strict sequence ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

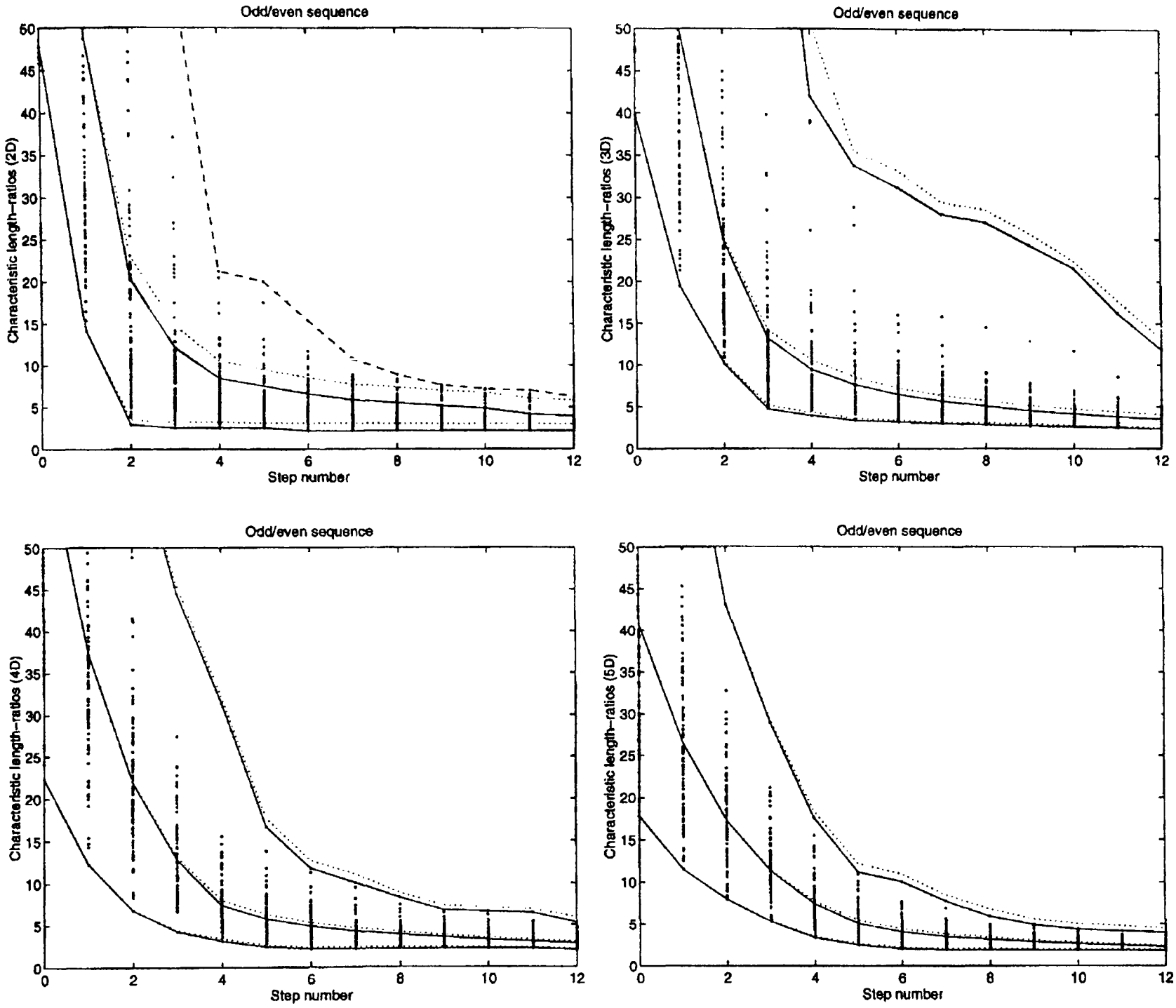


Figure 4.8: Characteristic length-ratios of odd/even sequence ellipsoids (noise uniformly distributed).

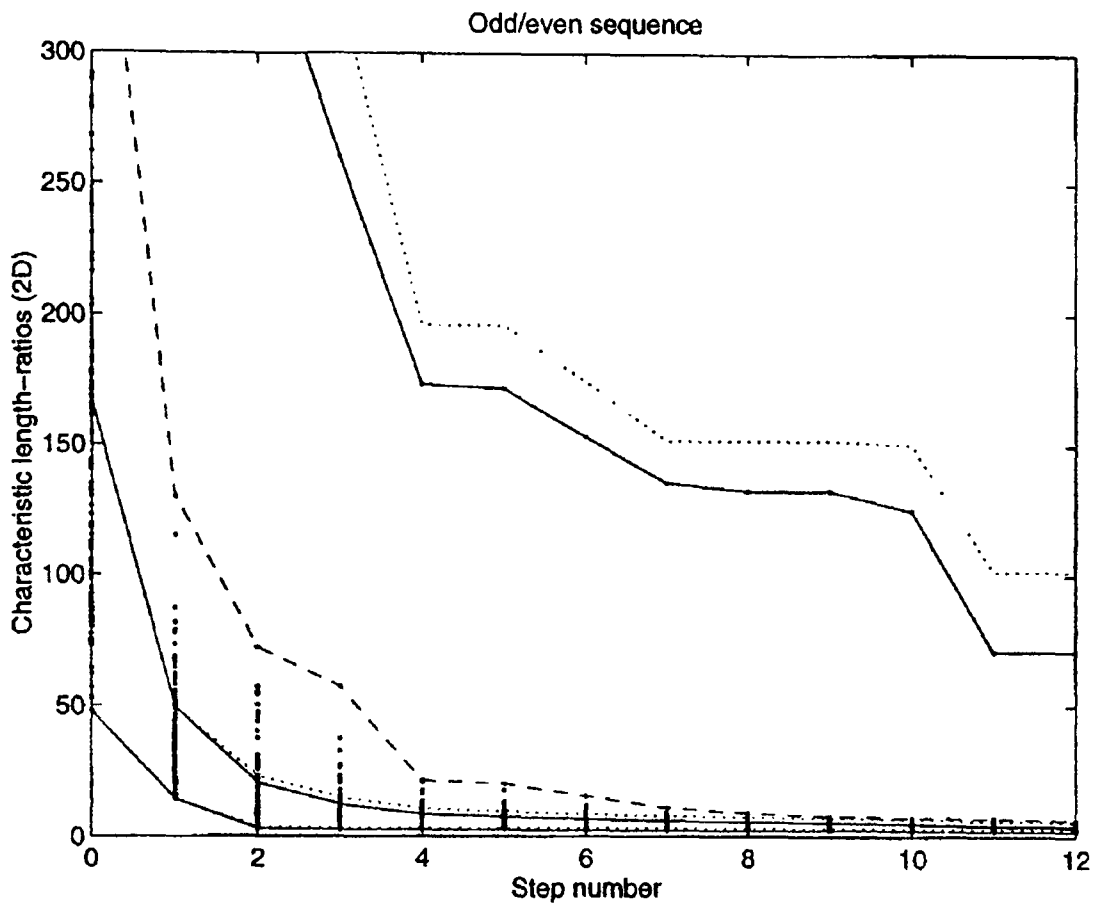


Figure 4.9: Characteristic length-ratios of odd/even sequence ellipsoids (noise uniformly distributed). (Showing the improvement for the worst case.)

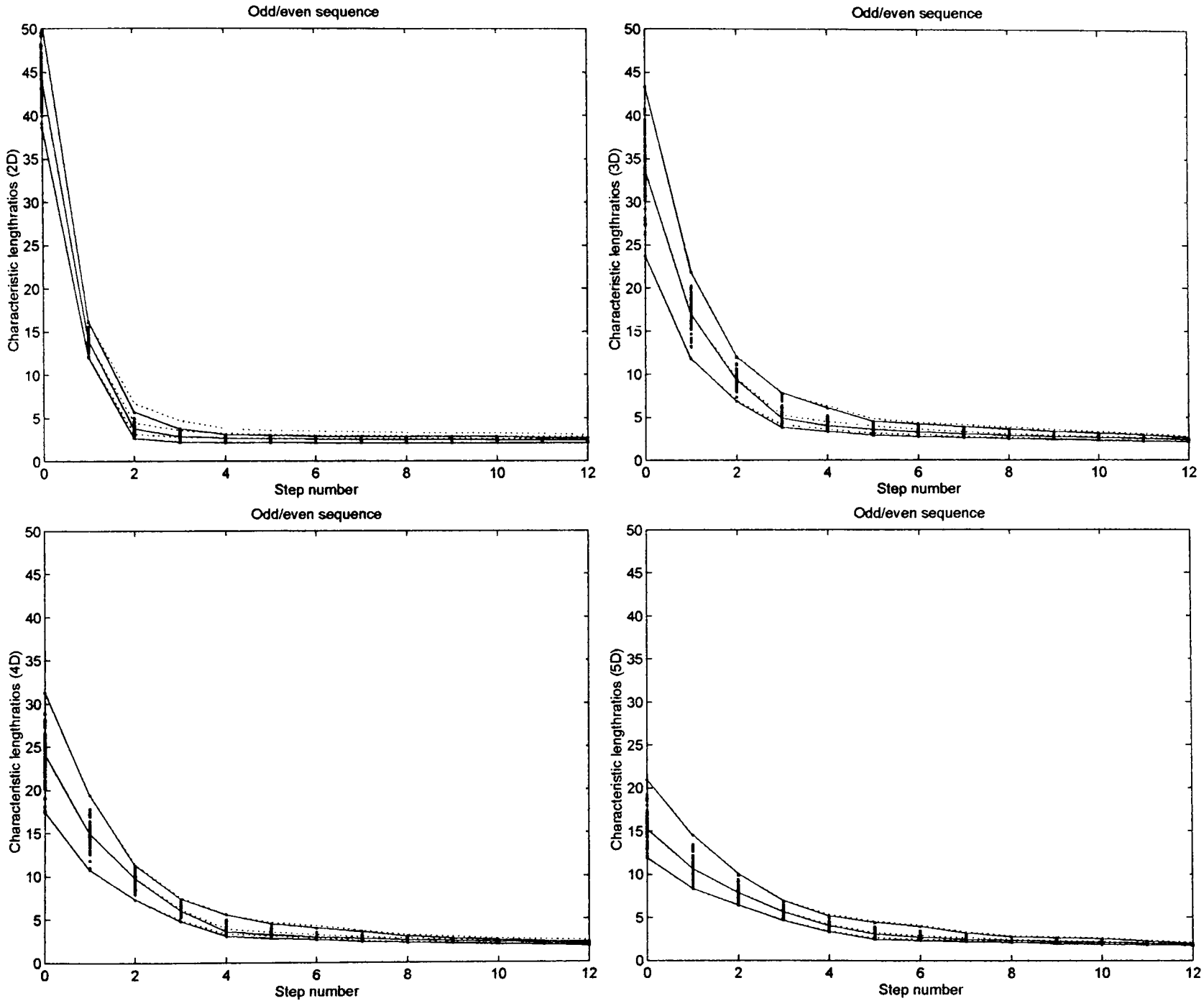


Figure 4.10: Characteristic length-ratios of odd/even sequence ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

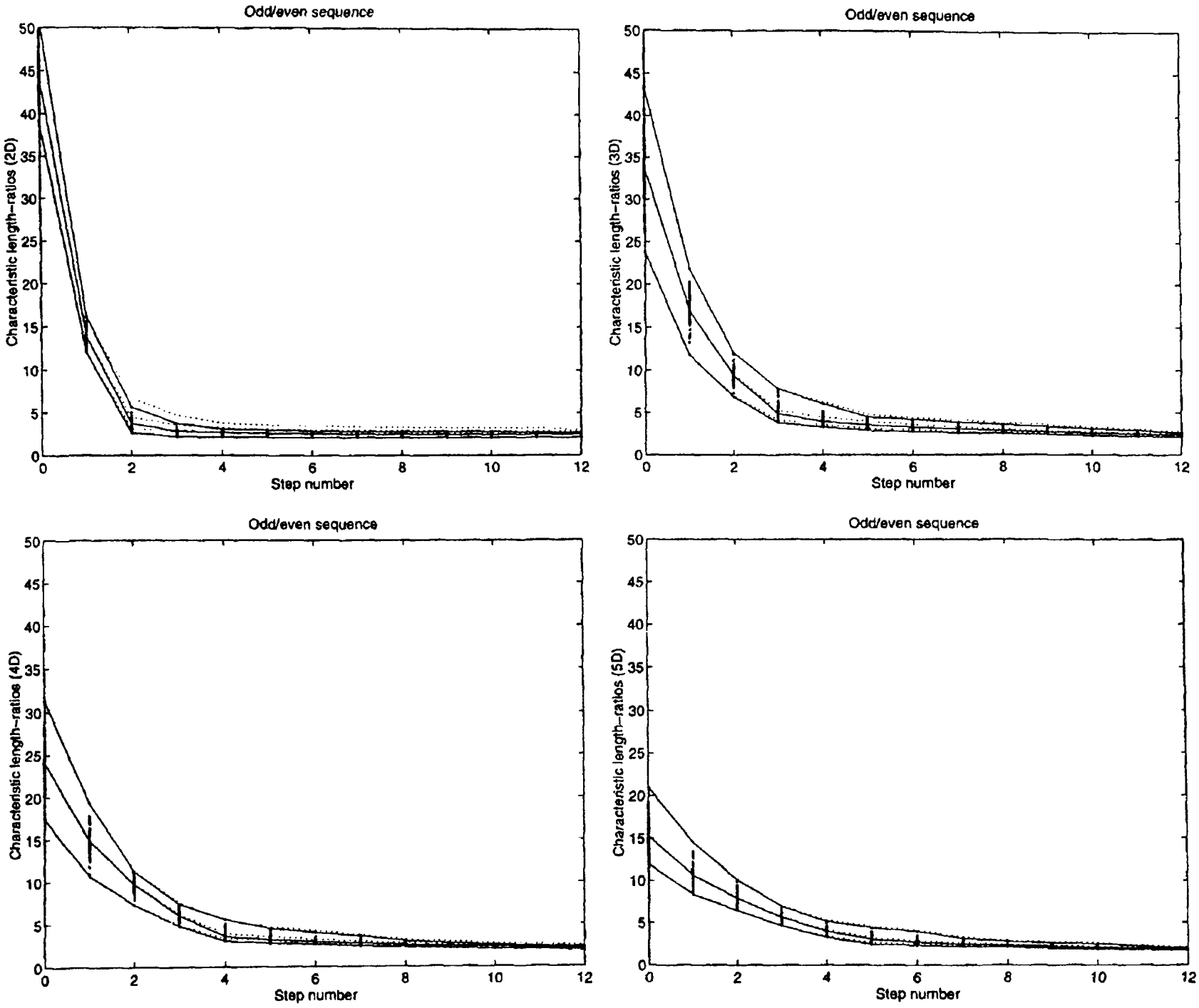


Figure 4.11: Characteristic length-ratios of "best second" ellipsoids (noise uniformly distributed).

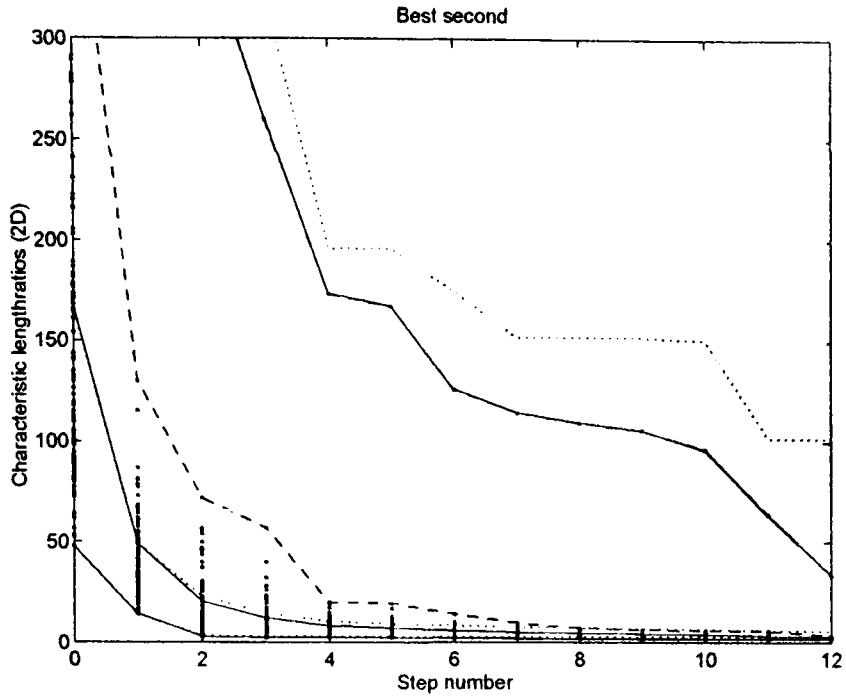


Figure 4.12: Characteristic length-ratios of "best second" ellipsoids (noise uniformly distributed). (Showing the improvement for the worst case.)

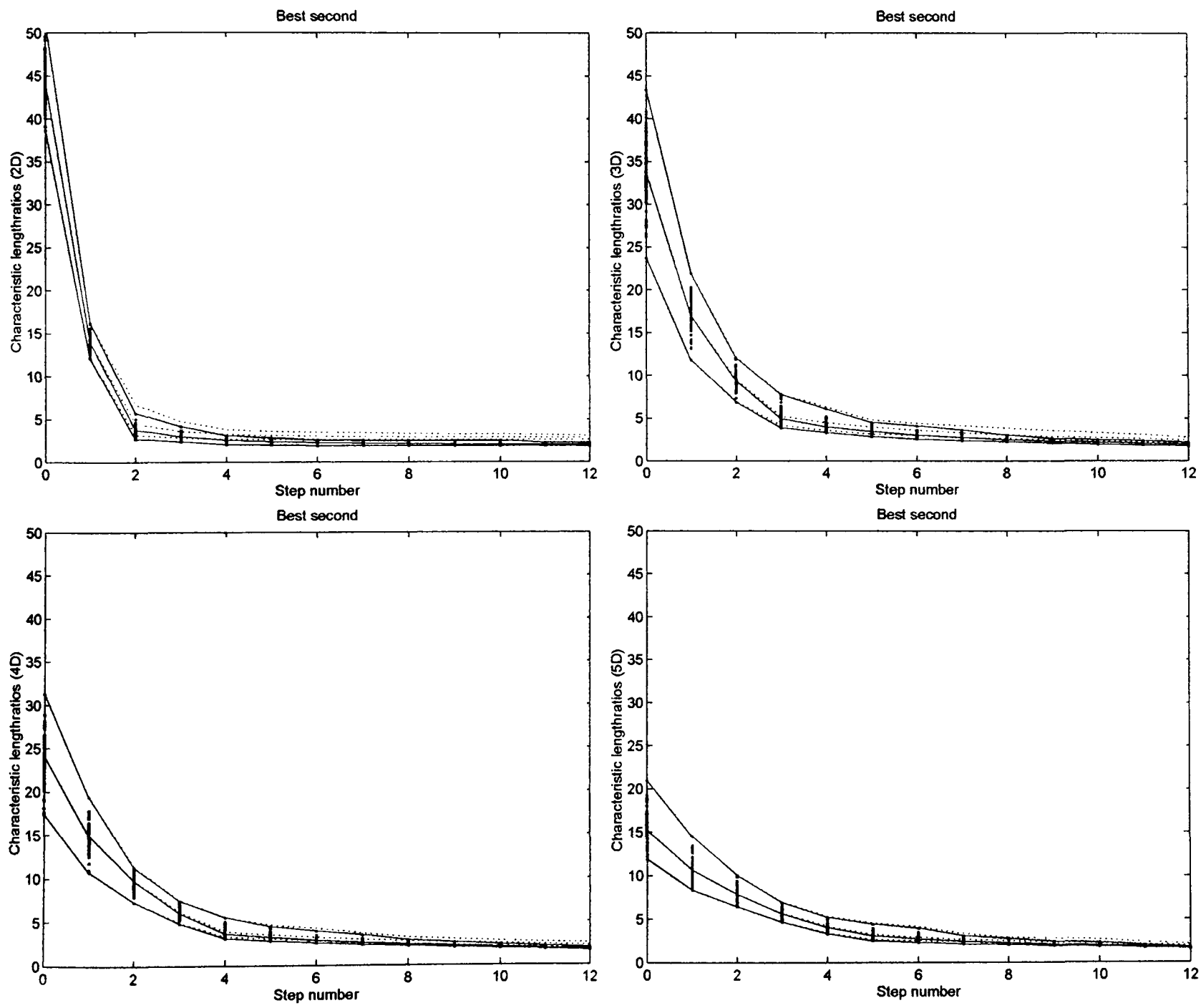


Figure 4.13: Characteristic length-ratios of "best second" ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

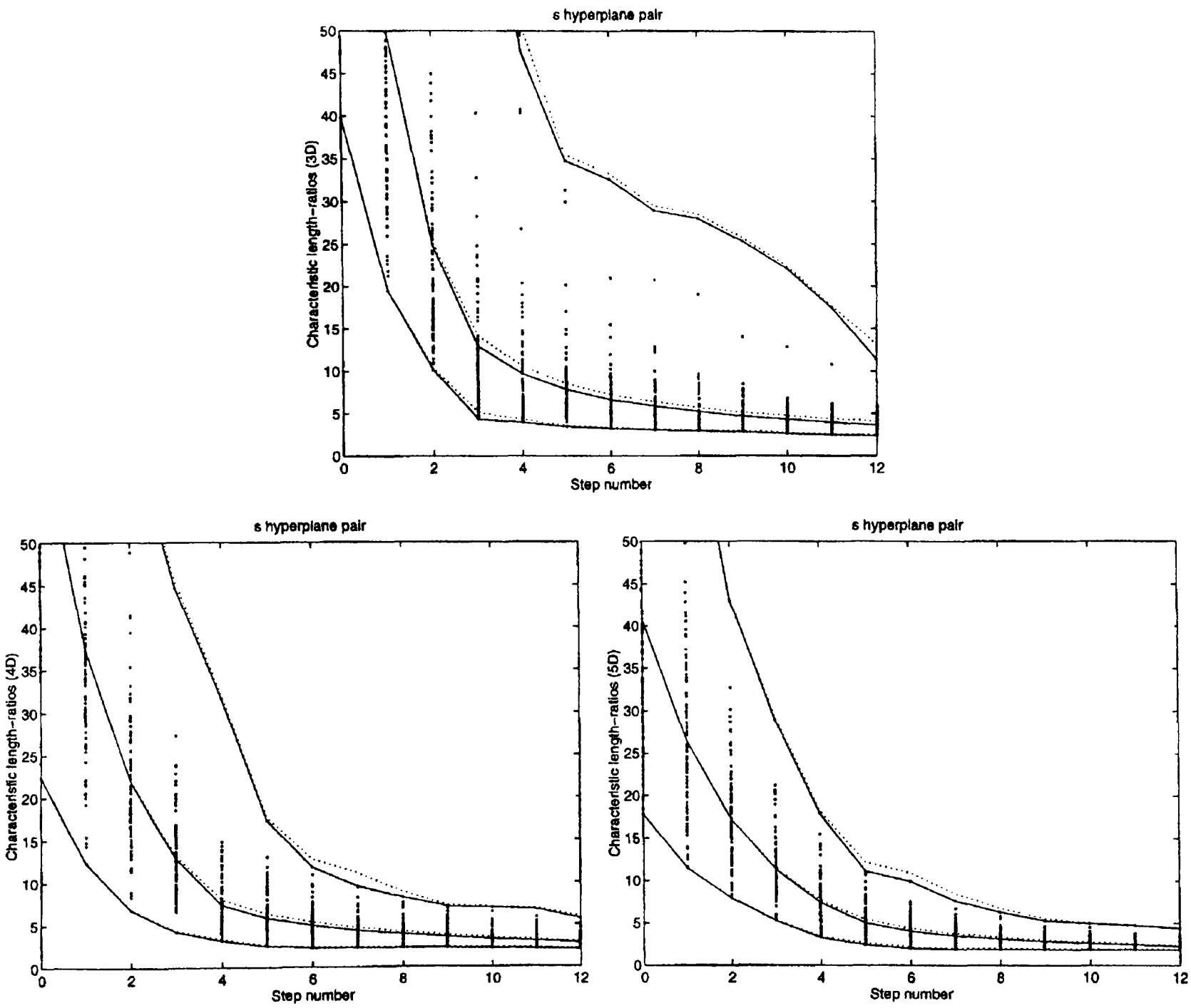


Figure 4.14: Characteristic length-ratios of s-hyperplane algorithm ellipsoids (noise uniformly distributed).

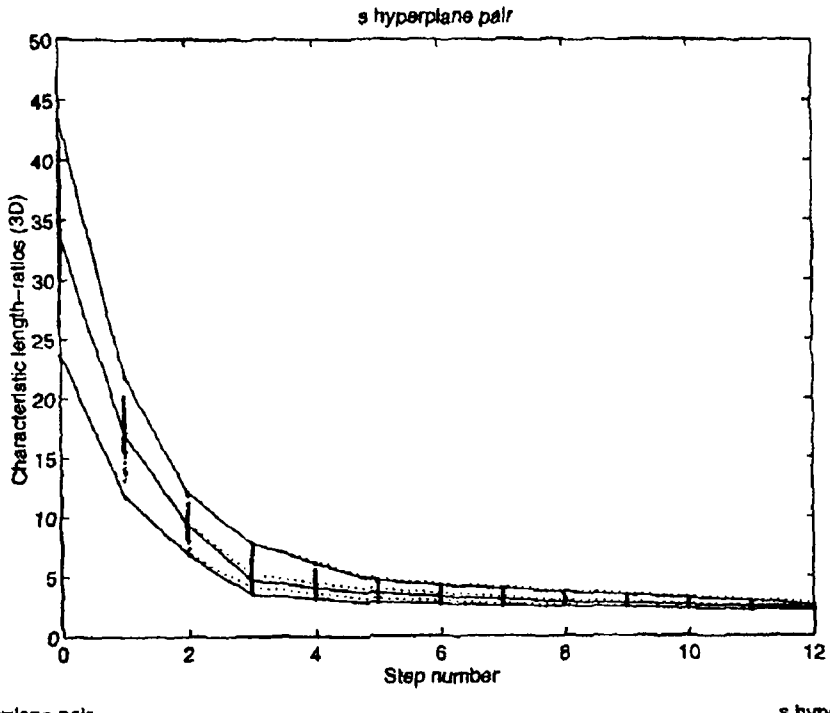


Figure 4.15: Characteristic length-ratios of s-hyperplane algorithm ellipsoids (noise with truncated normal distribution, $\sigma_t = 1/4\sqrt{3}$).

Chapter 5

Behrend-Löwner/John Ellipsoids for Two Observations at Once

The idea of the previous chapter was to iterate the process of finding the minimum-volume ellipsoid containing the intersection of an ellipsoid $\mathcal{E}_{k-1}$ with the region $\Pi_\ell \cap \Pi_k$ in parameter space consistent with two observations, where the minimum was taken over a(n) (effectively) two parameter family of ellipsoids. But, by the Behrend-Löwner/John theorem, there exists a unique minimum-volume ellipsoid $\mathcal{E}_k$ containing $\mathcal{E}_{k-1} \cap \Pi_\ell \cap \Pi_k$. In the present chapter this unique minimum-volume ellipsoid will be investigated.

Changing the origin of the subscripts, let the first ellipsoid be $\mathcal{E}_0 = \mathcal{E}(a_0, Q_0)$, let the strips be $\Pi_i = \Pi(n_i, y_i)$, $i = 1, 2$ and let the desired minimum-volume ellipsoid about $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ be $\mathcal{E}_2 = \mathcal{E}(a_2, Q_2)$, where, of course, the Q_i, $i = 0, 2$ are symmetric, positive-definite matrices.

Since Q_0 is symmetric and positive-definite, there exists a nonsingular matrix P such that $P^{\mathrm{T}} Q_0^{-1} P = I$. If x is set equal to $a_0 + Pz$, then

$$x \in \mathcal{E}(a_0, Q_0) \quad \Leftrightarrow \quad z \in \mathcal{E}(0, I), \tag{5.1}$$

$$x \in \Pi(n_i, y_i) \quad \Leftrightarrow \quad z \in \Pi(m_i, \nu_i), i = 1, 2, \tag{5.2}$$

where $m_i = P^{\mathrm{T}} n_i$, $\nu_i = y_i - {n_i}^{\mathrm{T}} a$, $i = 1, 2$. As ${m_i}^{\mathrm{T}} m_i = {n_i}^{\mathrm{T}} P P^{\mathrm{T}} n_i = {n_i}^{\mathrm{T}} Q_0 n_i$, ${m_i}^{\mathrm{T}} m_i = g_i$, $i = 1, 2$ (in line with the definitions of previous chapters).

But there also exists an orthogonal matrix O_1 such that ${O_1}^{\mathrm{T}} m_1 = |m_1| e_1 = \sqrt{g_1} e_1$ and ${O_1}^{\mathrm{T}} m_2 = n_{01} e_1 + n_{02} e_2$, where the e_k are canonical basis vectors for $\mathbb{R}^p$ and the n_{0i} can be chosen such that $n_{01} = h/\sqrt{g_1}$ and $n_{02} = \sqrt{g_1 g_2 - h^2}/\sqrt{g_1} > 0$ (as the $\mathbb{H}_1^{\pm}$ not parallel to the $\mathbb{H}_2^{\pm}$), so the required relations ${m_1}^{\mathrm{T}} O_1 {O_1}^{\mathrm{T}} m_2 = {m_1}^{\mathrm{T}} m_2 = {n_1}^{\mathrm{T}} P P^{\mathrm{T}} n_2 = {n_1}^{\mathrm{T}} Q_0 n_2 = h$ and ${m_2}^{\mathrm{T}} O_1 {O_1}^{\mathrm{T}} m_2 =$

$m_2{}^{\mathrm{T}} m_2 = n_2{}^{\mathrm{T}} Q_0 n_2 = n_2{}^{\mathrm{T}} P P^{\mathrm{T}} n_2 = h$ hold (where h is also defined in line with previous chapters). Hence, if $z = O_1 w$,

$$x \in \mathcal{E}(a_0, Q_0) \quad \Leftrightarrow \quad w \in \mathcal{E}(0, I), \tag{5.3}$$

$$x \in \Pi(n_1, y_1) \quad \Leftrightarrow \quad w \in \Pi(\sqrt{g_1} e_1, \nu_1), \tag{5.4}$$

$$x \in \Pi(n_2, y_2) \quad \Leftrightarrow \quad w \in \Pi(n_{01} e_1 + n_{02} e_2, \nu_2). \tag{5.5}$$

The order of volumes is preserved under the affine transformation which takes x to w, so the inverse image of the minimum-volume ellipsoid, $\tilde{\mathcal{E}}_2 = \mathcal{E}(\tilde{a}_2, \tilde{Q}_2)$, about $\mathcal{E}(0, I) \cap \Pi(\sqrt{g_1} e_1, \nu_1) \cap \Pi(n_{01} e_1 + n_{02} e_2, \nu_2)$ is the minimum-volume ellipsoid, $\mathcal{E}_2 = \mathcal{E}(a_2, Q_2)$, about $\mathcal{E}(a_0, Q_0) \cap \Pi(n_1, y_1) \cap \Pi(n_2, y_2)$. By symmetry, $\tilde{a}_1 = \xi_0 e_1 + \eta_0 e_2$ and $\tilde{Q}_2^{-1} = \alpha e_1 e_1{}^{\mathrm{T}} + \gamma(e_1 e_2{}^{\mathrm{T}} + e_2 e_1{}^{\mathrm{T}}) + \beta e_2 e_2{}^{\mathrm{T}} + \delta \sum_{i=1}^{p-2} e_{i+2} e_{i+2}{}^{\mathrm{T}}$.

Assume that the above transformations have already been made, so the problem becomes that of finding the values $a_2 = (\xi_0, \eta_0, 0^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$, α, β, γ and δ, such that

$$\begin{aligned} \mathcal{E}(a_2, Q_2^{-1}) &= \mathcal{E}\left(\begin{bmatrix} \xi_0 \\ \eta_0 \\ 0 \end{bmatrix}, \begin{bmatrix} \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix}^{-1} & 0 \\ 0 & \delta^{-1} I_{p-2} \end{bmatrix} \right) \\ &= \mathcal{E}\left(\begin{bmatrix} \bar{a}_2 \\ 0 \end{bmatrix}, \begin{bmatrix} \bar{Q}_2^{-1} & 0 \\ 0 & \delta^{-1} \mathrm{I}_{p-2} \end{bmatrix} \right) \\ &\supset \mathcal{E}(0, I) \cap \Pi(\sqrt{g_1} e_1, \nu_1) \cap \Pi(n_{01} e_1 + n_{02} e_2, \nu_2), \end{aligned} \tag{5.6}$$

and $\det Q_2$ is minimised (where I_{p-2} is the identity in $\mathbb{R}^{p-2}$). This last condition is, of course, equivalent to maximising $\det Q_2^{-1} = \delta^{(p-2)}(\alpha\beta - \gamma^2)$ (where p is the dimension of the parameter space, as before).

A change of notation is in order here. The strip $\bar{\Pi}(n, a, b)$ is defined by $x \in \bar{\Pi}(n, a, b) \Leftrightarrow a \leq n^{\mathrm{T}} x \leq b$, so $\Pi_1 = \Pi(\sqrt{g_1} e_1, \nu_1) = \bar{\Pi}(e_1, (\nu_1 - 1)/\sqrt{g_1}, (\nu_1 + 1)/\sqrt{g_1}) = \bar{\Pi}(e_1, \xi_1, \xi_2)$, and $\Pi_2 = \Pi(n_{01} e_1 + n_{02} e_2, \nu_2) = \bar{\Pi}(n_0, (\nu_2 - 1)/\sqrt{g_2}, (\nu_2 + 1)/\sqrt{g_2}) = \bar{\Pi}(n_0, \zeta_2, \zeta_2)$, where $\xi_{1,2} = (\nu_1 \mp 1)/\sqrt{g_1}$, $\zeta_{1,2} = (\nu_2 \mp 1)/\sqrt{g_2}$ and the unit vector $n_0 = (n_{01} e_1 + n_{02} e_2)/\sqrt{g_2}$.

The "missing" ellipsoid $\mathcal{E}_1 = \mathcal{E}(a_1, Q_1)$ can now be defined as being the minimum-volume ellipsoid about $\mathcal{E}_0(a_0, Q_0) \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$, where it can be assumed without loss of generality that $\xi_1 \xi_2 \geq \zeta_1 \zeta_2$, so that, if the minimum-volume ellipsoid about $\mathcal{E}_0 \cap \Pi_2$ is $\mathcal{E}_0$ (i.e., $\zeta_1 \zeta_2 \leq -1/p$), it is not necessarily the case that $\mathcal{E}_1 = \mathcal{E}_0$.

The case $p = 2$ will be considered later. For the moment, it is assumed that $p > 2$.

5.1 General Theorems Concerning $\mathcal{E}_2$

It is necessary to cite two results from Pronzato and Walter[23] here. The first is:

Lemma 5.1: Let $\mathcal{K}$ be a compact set of $\mathbb{R}^p$ and $\mathcal{E}(\mathcal{K})$ be the minimum-volume ellipsoid about $\mathcal{K}$. Then

$$\mathcal{E}(\mathcal{K}) = \left\{x \in \mathbb{R}^p : (x - c(w))^{\mathrm{T}} Q(w)^{-1} (x - c(w)) \leq 1\right\} \tag{5.7}$$

where

$$\begin{aligned} Q(w) &= p(M(w) - c(w)c(w)^{\mathrm{T}}) && (5.8)\\ M(w) &= \int_{\mathcal{K}} xx^{\mathrm{T}} w(\mathrm{d}x) && (5.9)\\ c(w) &= \int_{\mathcal{K}} x w(\mathrm{d}x) && (5.10) \end{aligned}$$

and

$$w = \arg\max_{w' \in \Omega} \ln\det(M(w') - c(w')c(w')^{\mathrm{T}}), \tag{5.11}$$

where Ω is the set of distributions w' over $\mathcal{K}$ such that

$$\int_{\mathcal{K}} w'(\mathrm{d}x) = 1 \tag{5.12}$$

□

and the second is

Lemma 5.2:

1. A distribution w supported by at most $\frac{1}{2}p(p+3)$ (and at least $p+1$) points of $\mathcal{K}$ always exists. These support points are located on the boundary of the complex closure of $\mathcal{K}$. When there are only $p+1$ support points, they are uniformly weighted and $c(w)$ corresponds to their centre of gravity.

2. w is not necessarily unique (although $M(w)$ and $c(w)$ are), but the set of all distributions satisfying equation (5.11) is convex.

3. $\forall x \in \mathcal{K}, d(w,x) \leq 0$, where

$$d(w',x) = x^{\mathrm{T}} M(w')^{-1} x + \frac{[x^{\mathrm{T}} M(w')^{-1} c(w') - 1]^2}{1 - c(w')^{\mathrm{T}} M(w')^{-1} c(w')} - p - 1 \tag{5.13}$$

4. $max_{x \in \mathcal{K}} d(w,x) = \min_{w' \in \Omega} \max_{x \in \mathcal{K}} d(w',x)$.

□

Definition 5.1: *A distribution satisfying equation (5.11) will be called* optimal *for* $\mathcal{K}$. □

A symmetry result can now be deduced:

Corollary 5.3: Suppose $\mathcal{K}$ is a compact set of $\mathbb{R}^p$ such that $\mathcal{K}$ is unaffected by rotations which leave e_1 and e_2 unchanged. Then there exists an optimal distribution which is unchanged by rotations which leave e_1 and e_2 unaltered. □

Proof 5.3: *By Lemma 5.2, there exists an optimal distribution* w_1 *for* $\mathcal{K}$. *This distribution is not necessarily symmetric under the interchange of* e_i *and* e_j, $i, j \notin \{1, 2\}$, *but the distribution* w_2 *derived from* w_2 *by such an interchange will also be optimal for* $\mathcal{K}$. *Similarly, a distribution* w_3 *derived from* w_1 *by a reflection leaving* e_1 *and* e_2 *unchanged will also be optimal. Suppose there are* $k-1$ *distinct such distributions derived from such interchanges or reflections:* $w_2, w_3, \ldots, w_k$, *say. Then the distribution* $w = (\sum_{i=1}^{k} w_i)/k$ *will also be optimal for* $\mathcal{K}$, *and* w *is unaffected by rotations not affecting* e_i *and* e_j. ■

Corollary 5.4: If $\mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2) \cap \bar{\Pi}(n_{01}e_1 + n_{02}e_2, \zeta_1, \zeta_2)$ contains the support points of an optimal distribution w for $\mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$, then $\mathcal{E}_2 = \mathcal{E}_1$. □

Proof 5.4: *Let* $\mathcal{K}_1 = \mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$, $\mathcal{K}_2 = \mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2) \cap \bar{\Pi}(n_{01}e_1 + n_{02}e_2, \zeta_1, \zeta_2)$. *Obviously,* $\mathcal{K}_i$ *is compact (and convex),* $i = 1, 2$. *Let* $\Omega(\mathcal{K}_i) = \left\{ w' : \int_{x \in \mathcal{K}_i} w'(dx) = 1 \right\}$, $i = 1, 2$, *so* $w \in \Omega(\mathcal{K}_1)$ *is optimal for* $\mathcal{K}_1$. *Implicitly assuming the obvious mapping between the distributions on* $\mathcal{K}_1$ *and those on its subset* $\mathcal{K}_2$, *it follows that* $w \in \Omega(\mathcal{K}_2)$, *as* w *vanishes except on* $\mathcal{K}_2$. *But* $\Omega(\mathcal{K}_1) \supset \Omega(\mathcal{K}_2)$, *and this implies that*

$$\begin{aligned} \ln\det\left(M(w) - c(w)c(w)^{\mathrm{T}}\right) &\leq \max_{w' \in \Omega(\mathcal{K}_2)} \ln\det\left(M(w') - c(w')c(w')^{\mathrm{T}}\right) \\ &\leq \max_{w' \in \Omega(\mathcal{K}_1)} \ln\det\left(M(w') - c(w')c(w')^{\mathrm{T}}\right), \end{aligned}$$

and the equality of the outer members of this chain of inequalities implies that w *is optimal for* $\mathcal{K}_2$ *as well. But, if two compact sets have the same optimal distribution, they are contained in the same minimum-volume ellipsoid, by Lemma 5.1.* ■

It will be useful to quote the results which define $\mathcal{E}_1$: if $\xi_1\xi_2 < -1/p$,

$$Q_1 = \mathrm{I}, \tag{5.14}$$

$$a_1 = 0; \tag{5.15}$$

if $\xi_1\xi_2 \geq -1/p$,

$$Q_1^{-1} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \beta_1 \mathrm{I}_{p-1} \end{bmatrix} \in \mathbb{R}^{p \times p} \tag{5.16}$$

$$a_1 = \left[\begin{array}{cc} \xi_{01} & 0 \end{array} \right]^{\mathrm{T}} \in \mathbb{R}^p, \tag{5.17}$$

where

$$\beta_1 = \frac{1 - \frac{\xi_1^2+\xi_2^2}{2} - \sqrt{\left(\frac{\xi_2^2-\xi_1^2}{2}\right)^2 + \frac{(1-\xi_1^2)(1-\xi_2^2)}{p^2}}}{(1-\xi_1^2)(1-\xi_2^2)} \tag{5.18}$$

$$\alpha_1 = \left(\frac{\sqrt{1-\beta_1(1-\xi_1^2)} + \sqrt{1-\beta_1(1-\xi_2^2)}}{\xi_2 - \xi_1} \right)^2 \tag{5.19}$$

$$\xi_{01} = \xi_1 + \sqrt{\frac{1-\beta_1(1-\xi_1^2)}{\alpha_1}}. \tag{5.20}$$

These results are the translation into the notation of the present text of a simple extension of König and Pallaschke's [16] results.

Equation (5.20) will first be put into a more symmetric form not involving α_1 or β_1.

Now

$$1 - \beta_1(1-\xi_1^2) = \frac{\sqrt{\left(\frac{\xi_2^2-\xi_1^2}{2}\right)^2 + \frac{(1-\xi_1^2)(1-\xi_2^2)}{p^2}} - \frac{\xi_2^2-\xi_1^2}{2}}{1-\xi_2}$$

and there is a similar expression for $1 - \beta_1(1-\xi_2^2)$. Together, these yield $(1-\beta_1(1-\xi_1^2))(1-\beta_1(1-\xi_2^2)) = 1/p^2$ and then

$$\begin{aligned} \xi_{01} &= \xi_1 + \frac{(\xi_2-\xi_1)\sqrt{1-\beta_1(1-\xi_1^2)}}{\sqrt{1-\beta_1(1-\xi_1^2)} + \sqrt{1-\beta_1(1-\xi_2^2)}} \\ &= \xi_1 + \frac{\sqrt{1-\beta_1(1-\xi_1^2)}\sqrt{1-\beta_1(1-\xi_2^2)} - 1 + \beta_1(1-\xi_1^2)}{\beta_1(\xi_1+\xi_2)} \\ &= \frac{1+\xi_1\xi_2}{\xi_1+\xi_2} - \frac{p-1}{p}\frac{1}{\beta_1(\xi_1+\xi_2)} \\ &= \frac{1+\xi_1\xi_2}{\xi_1+\xi_2} - \frac{p\left[1 - \frac{\xi_1^2+\xi_2^2}{2} + \sqrt{\left(\frac{\xi_2^2-\xi_1^2}{2}\right)^2 + \frac{(1-\xi_1^2)(1-\xi_2^2)}{p^2}}\right]}{(p+1)(\xi_1+\xi_2)} \end{aligned} \tag{5.21}$$

Equations (5.18), (5.19) and (5.21) will now be used to find a family of optimal distributions for $\mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$.

Theorem 5.5: Let $x_{ijkl} = \xi_i e_1 + \eta_{ij} e_2 + (-1)^k \sqrt{1-\xi_i^2-\eta_{ij}^2}\, e_{l+2}$, $i,j,k = 1,2$, $l = 1,\ldots,(p-2)$, for some η_{ij} such that $\eta_{ij} \in [-\sqrt{1-\xi_i^2}, \sqrt{1-\xi_i^2}]$ $\forall i,j = 1,2$ (so that $x_{ijkl} \in \mathcal{E}_0 \cap \bar{\Pi}(e_1,\xi_1,\xi_2)$ $\forall i,j,k = 1,2, l = 1,\ldots,p-1$) be the support points of a distribution w which takes on the nonnegative values w_{ijkl} at these support points.

Then, if

$$\eta_{11} \le \eta_{12} \quad \text{and} \quad \eta_{21} \le \eta_{22}, \tag{5.22}$$

the η's also satisfy

$$-(\xi_2 - \xi_{01})\eta_{11}\eta_{12} - (\xi_{01} - \xi_1)\eta_{21}\eta_{22} = \frac{1 + \xi_1\xi_2 - (\xi_1 + \xi_2)\xi_{01}}{p-1}(\xi_2 - \xi_1) \tag{5.23}$$

and the weights w_{ijkl} are given by

$$\begin{aligned}
w_{ij11} &= \frac{(\xi_2 - \xi_{01})\eta_{12}}{2(p-2)(\xi_2 - \xi_1)(\eta_{12} - \eta_{11})} && (5.24)\\
w_{ij12} &= \frac{-(\xi_2 - \xi_{01})\eta_{11}}{2(p-2)(\xi_2 - \xi_1)(\eta_{12} - \eta_{11})} && (5.25)\\
w_{ij21} &= \frac{(\xi_{01} - \xi_1)\eta_{22}}{2(p-2)(\xi_2 - \xi_1)(\eta_{22} - \eta_{21})} && (5.26)\\
w_{ij22} &= \frac{-(\xi_{01} - \xi_1)\eta_{21}}{2(p-2)(\xi_2 - \xi_1)(\eta_{22} - \eta_{21})}, \ \forall i, j = 1, 2, && (5.27)
\end{aligned}$$

the distribution w is optimal for $\mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$. □

Proof 5.5:

Consider the distribution w with support points $x_{ijkl} = \xi_i e_1 + \eta_{ij} e_2 + (-1)^k \sqrt{1 - \xi_i^2 - \eta_{ij}^2} e_{l+2}$, $i, j, k = 1, 2$, $l = 1, \ldots, (p-2)$, for some η_{ij} such that $\eta_{ij} \in [-\sqrt{1-\xi_i^2}, \sqrt{1-\xi_i^2}]$ $\forall i, j = 1, 2$, and with the symmetric values $w_{ijkl} = w_{ij}$ for some nonnegative w_{ij} and $k = 1, 2, l = 1, \ldots, p-2$. Then the condition that the distribution w be normalised is that

$$\sum_{ijkl} w_{ijkl} = 2(p-2) \sum_{ij} w_{ij} = 1. \tag{5.28}$$

The vector $c(w)$ is given by

$$\begin{aligned}
c(w) &= \sum_{ijkl} w_{ijkl} x_{ijkl} \\
&= 2(p-2) \left[\left(\sum_{ij} w_{ij} \xi_i \right) e_1 + \left(\sum_{ij} w_{ij} \eta_{ij} \right) e_2 \right] && (5.29)
\end{aligned}$$

and the matrix $M(w)$ is given by

$$\begin{aligned}
M(w) &= \sum_{ijkl} w_{ijkl} x_{ijkl} {x_{ijkl}}^{\mathrm{T}} \\
&= 2(p-2) \left[\left(\sum_{ij} w_{ij} \xi_i^2 \right) e_1 {e_1}^{\mathrm{T}} \right. \\
&\qquad \left. + \left(\sum_{ij} w_{ij} \xi_i \eta_{ij} \right) (e_1 {e_2}^{\mathrm{T}} + e_2 {e_1}^{\mathrm{T}}) + \left(\sum_{ij} w_{ij} \eta_{ij}^2 \right) e_2 {e_2}^{\mathrm{T}} \right] \\
&\quad + 2 \left(\sum_{ij} w_{ij} (1 - \xi_i^2 - \eta_{ij}^2) \right) \left(\sum_{l} e_{l+2} {e_{l+2}}^{\mathrm{T}} \right). && (5.30)
\end{aligned}$$

Now, if it were to be the case that

$$c(w) = a_1 \text{ and} \tag{5.31}$$

$$M(w) = c(w)c(w)^{\mathrm{T}} + \frac{1}{p}Q_1, \tag{5.32}$$

it would also be the case that

$$\begin{aligned}
x \in \mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2) &\Rightarrow x \in \mathcal{E}(a_1, Q_1) \\
&\Leftrightarrow (x - a_1)^{\mathrm{T}} Q^{-1} (x - a_1) \leq 1 \\
&\Leftrightarrow (x - c(w))^{\mathrm{T}} \\
&\quad \times \left(M(w) - c(w)c(w)^{\mathrm{T}}\right)^{-1} (x - c(w)) \leq p \\
&\Leftrightarrow (x - c(w))^{\mathrm{T}} M(w)^{-1} \\
&\quad \times \left[\mathrm{I}_p + \frac{c(w)c(w)^{\mathrm{T}} M(w)^{-1}}{1 - c(w)^{\mathrm{T}} M(w)^{-1} c(w)}\right] (x - c(w)) \leq p \\
&\Leftrightarrow d(x, w) \leq 0,
\end{aligned}$$

(where the penultimate equivalence follows from the matrix inversion lemma), that is, w would be optimal for $\mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$, by Lemma 5.2.

But, if the w_{ijkl} are given by equations (5.24) to (5.27),

$$c(w) = a_1, \tag{5.33}$$

$$M(w) = \begin{bmatrix} M_{11}(w) & 0 & 0 \\ 0 & M_{22}(w) & 0 \\ 0 & 0 & \bar{M}(w)\mathrm{I}_{p-2} \end{bmatrix}, \tag{5.34}$$

where

$$M_{11}(w) = (\xi_1 + \xi_2)\xi_{01} - \xi_1\xi_2, \tag{5.35}$$

$$M_{22}(w) = -\frac{(\xi_2 - \xi_{01})\eta_{11}\eta_{12} + (\xi_{01} - \xi_1)\eta_{21}\eta_{22}}{\xi_2 - \xi_1}, \tag{5.36}$$

$$\bar{M}(w) = \frac{1}{p-2}(1 - M_{11}(w) - M_{22}(w)). \tag{5.37}$$

But then $M_{11}(w) - \xi_{01}^2 = (\xi_{01} - \xi_1)(\xi_2 - \xi_{01}) = p^{-1}\alpha_1^{-1}$, $M_{22} = (p-1)^{-1}[1 + \xi_1\xi_2 - (\xi_1 + \xi_2)\xi_{01}] = p^{-1}\beta_1^{-1}$ and $\bar{M}(w) = (p-2)^{-1}(1 + \xi_{01}^2 - p^{-1}\alpha_1^{-1} - p^{-1}\beta_1^{-1}) = p^{-1}\beta_1^{-1}$. ■

A relation between α, β, γ, δ, ξ_0 and η_0 can be deduced from methods like the above.

Theorem 5.6: If $p > 2$, the components of Q^{-1} and c obey

$$\delta = \frac{(p-2)(\alpha\beta - \gamma^2)}{p(1-\xi_0^2-\eta_0^2)(\alpha\beta-\gamma^2) - \alpha - \beta}. \tag{5.38}$$

□

Proof 5.6: *By Pronzato and Walter [23], the sequence of distributions* $\{w^{(i)}\}$ *converges to an optimal distribution for* $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, *where* $w^{(i)}$ *is given by choosing*

$$x_i' = \arg\max\{d(w^{(i)}, x) : x \in \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2\} \tag{5.39}$$

and setting

$$w^{(i+1)} = (1-s_i)w^{(i)} + s_i\tilde{x}_i', \tag{5.40}$$

where $\tilde{x}_i'$ *is the distribution concentrated at* x_i',

$$d(w, x) = x^{\mathrm{T}} M(w)^{-1} x + \frac{(1 - x^{\mathrm{T}} M(w)^{-1} c(w))^2}{1 - c(w)^{\mathrm{T}} M(w)^{-1} c(w)} - p - 1 \tag{5.41}$$

and

$$s_i = \frac{d(w^{(i)}, x_i')}{(p + d(w^{(i)}, x_i'))(p+1)}. \tag{5.42}$$

The sequence is started by choosing support points $x_1^{(0)}, \ldots, x_{k_0}^{(0)} \in \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$ *such that* $k_0^{-1}\sum_{i=1}^{k_0} x_i^{(0)} x_i^{(0)\mathrm{T}} - k_0^{-2}\left(\sum_{i=1}^{k_0} x_i^{(0)}\right)\left(\sum_{i=1}^{k_0} x_i^{(0)}\right)^{\mathrm{T}}$ *is nonsingular, and letting the weights* $w_1^{(0)} = \cdots = w_{k_0}^{(0)} = 1/k_0$.

In particular, let $\{w_1^{(i)}\}$ *be the convergent sequence started by making a definite choice of* $x_1^{(0)}, \ldots, x_{k_0}^{(0)} \in \partial\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ *and a definite choice of* x_i' *at each step* i. *Then* $\{w_1^{(i)}\}$ *converges, to* w_1, *say. Now*

$$d(w, x) = (x - c(w))^{\mathrm{T}}(M(w) - c(w)c(w)^{\mathrm{T}})^{-1}(x - c(w)) - p - 1, \tag{5.43}$$

has no maximum on any unbounded set in $\mathbb{R}^p$, *provided* $M(w) - c(w)c(w)^{\mathrm{T}}$ *is positive definite. Hence, if* $M(w) - c(w)c(w)^{\mathrm{T}}$ *is positive definite, any maximum of* $d(w, x)$ *over* $\partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$ *cannot be on the unbounded sets* $\mathbb{H}_{1,2}^{\pm}$ *or* $\mathbb{H}_1^{\pm} \cap \mathbb{H}_2^{\pm'}$ *unless it is also in* $\partial\mathcal{E}_0$. *But, if* $M(w_1^{(0)}) - c(w_1^{(0)})c(w_1^{(0)})^{\mathrm{T}}$ *is nonsingular, it is positive definite, and, if* $M(w_1^{(i)}) - c(w_1^{(i)})c(w_1^{(i)})^{\mathrm{T}}$ *is positive definite, then* $M(w_1^{(i+1)}) - c(w_1^{(i+1)})c(w_1^{(i+1)})^{\mathrm{T}}$ *is positive definite by choice of* s_i. *Thus,* $x_i' \in \partial\mathcal{E}_0$ *for each* i, *and so the set of support points of* $w_1^{(i)}$ *is contained in the closed set* $\partial\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ *for each* i. *Consequently, the set of support points of* w_1 *is contained in* $\partial\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$.

Let w *be the symmetric density derived from* w_1 *by the process described in the proof of Corollary 5.3. The set of support points of* w *will also be contained in* $\partial\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$.

Then

$$\begin{aligned} M(w) &= \int_{\mathcal{E}_0\cap\Pi_1\cap\Pi_2} xx^{\mathrm{T}}w(\mathrm{d}x) \\ &= \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} \left(\sum_{j=1}^{p}({e_j}^{\mathrm{T}}x)e_j\right)\left(\sum_{j=1}^{p}({e_j}^{\mathrm{T}}x)e_j\right)^{\mathrm{T}}w(\mathrm{d}x) \\ &= \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_1}^{\mathrm{T}}x)^2 w(\mathrm{d}x)e_1{e_1}^{\mathrm{T}} + \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_1}^{\mathrm{T}}x)({e_2}^{\mathrm{T}}x)w(\mathrm{d}x)(e_1{e_2}^{\mathrm{T}} + e_2{e_1}^{\mathrm{T}}) + \\ &\qquad \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_2}^{\mathrm{T}}x)^2 w(\mathrm{d}x)e_2{e_2}^{\mathrm{T}} + \sum_{j=1}^{p-2}\int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_{j+2}}^{\mathrm{T}}x)^2 w(\mathrm{d}x)e_{j+2}{e_{j+2}}^{\mathrm{T}}, \end{aligned} \tag{5.44}$$

by the symmetry under reflections. But clearly

$$\sum_{j=1}^{p-2}\int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_{j+2}}^{\mathrm{T}}x)^2 w(\mathrm{d}x) = \sum_{j=1}^{p-2}\int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} \left(1 - ({e_1}^{\mathrm{T}}x)^2 - ({e_2}^{\mathrm{T}}x)^2\right) w(\mathrm{d}x) \tag{5.45}$$

and also

$$\sum_{j=1}^{p-2}\int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_{j+2}}^{\mathrm{T}}x)^2 w(\mathrm{d}x) = (p-2)\int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_{k+2}}^{\mathrm{T}}x)^2 w(\mathrm{d}x),\ k = 1,\ldots,p-2, \tag{5.46}$$

by symmetry, so

$$\begin{aligned} M(w) &= \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_1}^{\mathrm{T}}x)^2 w(\mathrm{d}x)e_1{e_1}^{\mathrm{T}} + \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_1}^{\mathrm{T}}x)({e_2}^{\mathrm{T}}x)w(\mathrm{d}x)(e_1{e_2}^{\mathrm{T}} + e_2{e_1}^{\mathrm{T}}) + \\ &\qquad \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_2}^{\mathrm{T}}x)^2 w(\mathrm{d}x)e_2{e_2}^{\mathrm{T}} + \\ &\qquad \frac{1}{p-2}\left(1 - \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_1}^{\mathrm{T}}x)^2 w(\mathrm{d}x) - \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} ({e_2}^{\mathrm{T}}x)^2 w(\mathrm{d}x)\right)\sum_{j=1}^{p-2} e_{j+2}{e_{j+2}}^{\mathrm{T}} \\ &= \begin{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{12} & M_{22} \end{bmatrix} & 0 \\ 0 & \frac{1-M_{11}-M_{22}}{p-2}\mathrm{I}_{p-2} \end{bmatrix}. \end{aligned} \tag{5.47}$$

Similarly,

$$c(w) = \int_{\partial\mathcal{E}_0\cap\Pi_1\cap\Pi_2} xw(\mathrm{d}x) = \begin{bmatrix} \xi_0 \\ \eta_0 \\ 0 \end{bmatrix}. \tag{5.48}$$

The relations between Q^{-1} and $M(w)$,

$$Q^{-1} = \begin{bmatrix} \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix} & 0 \\ 0 & \delta\mathrm{I}_{p-2} \end{bmatrix} = p^{-1}(M(w) - c(w)c(w)^{\mathrm{T}})^{-1}, \tag{5.49}$$

then allow equation (5.38) to be deduced. ■

It is now desirable to find conditions such that $\mathcal{E}_2 = \mathcal{E}_1$.

5.2 Conditions for $\mathcal{E}_2 = \mathcal{E}_0$

The treatment will have to be divided into two cases here: $\mathcal{E}_1 \neq \mathcal{E}_0$ and $\mathcal{E}_1 = \mathcal{E}_0$.

5.2.1 $\mathcal{E}_1 \neq \mathcal{E}(0, I)$

Here it must be the case that $\xi_1 \xi_2 > -1/p$.

Some sets need to be defined here: $S = \{(\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in [-\eta_{110}, 0] \times [0, \eta_{120}] \times [-\eta_{210}, 0] \times [0, \eta_{220}] : a\eta_{11}\eta_{12} + b\eta_{21}\eta_{22} + c = 0\}$ and $T = \{(\eta_{11}, \eta_{21}) \in [-\eta_{110}, 0] \times [-\eta_{210}, 0] : \exists \eta_{12}, \eta_{22} : (\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in S\}$, where $a, b, c > 0$ and $\eta_{110}, \eta_{120}, \eta_{210}, \eta_{220} \geq 0$.

These sets have the properties given by the Lemmas below.

Lemma 5.7:

$$S = S_{01} \cup S_{10} \cup S_{11}, \tag{5.50}$$

where

$$S_{01} = \begin{cases} \emptyset, & \text{if } b\eta_{210}\eta_{220} < c; \\ \{(0, \eta_{12}, \eta_{21}, -c/b\eta_{21}) : & \\ \qquad \eta_{12} \in [0, \eta_{120}], \eta_{21} \in [-\eta_{210}, -c/b\eta_{220}]\}, & \text{if } b\eta_{210}\eta_{220} \geq c, \end{cases} \tag{5.51}$$

$$S_{10} = \begin{cases} \emptyset, & \text{if } a\eta_{110}\eta_{120} < c; \\ \{(\eta_{110}, -c/a\eta_{11}, 0, \eta_{22}) : & \\ \qquad \eta_{11} \in [-\eta_{110}, -c/a\eta_{120}], \eta_{22} \in [0, \eta_{220}]\}, & \text{if } a\eta_{110}\eta_{120} \geq c, \end{cases} \tag{5.52}$$

and

$$
S_{11} = \begin{cases}
\emptyset, & \text{if } a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} < c; \\
\{(\eta_{11}, 0, \eta_{21}, -c/b\eta_{21}) : & \\
\quad \eta_{11} \in [-\eta_{110}, 0), \eta_{21} \in [-\eta_{210}, -c/b\eta_{220}]\}, & \text{if } \eta_{120} = 0 \text{ and } b\eta_{210}\eta_{220} \geq c; \\
\{(\eta_{11}, -c/a\eta_{11}, \eta_{21}, 0) : & \\
\quad \eta_{11} \in [-\eta_{110}, -c/a\eta_{120}], \eta_{21} \in [-\eta_{210}, 0)\}, & \text{if } \eta_{220} = 0 \text{ and } a\eta_{110}\eta_{120} \geq c; \\
\{(\eta_{11}, \eta_{12}, \eta_{21}, -(a\eta_{11}\eta_{12} + c)/b\eta_{21}) : & \\
\quad \eta_{11} \in [-\eta_{110}, 0) \cap & \\
\qquad (-\infty, (b\eta_{210}\eta_{220} - c)/a\eta_{120}]\}, & \\
\quad \eta_{21} \in [-\eta_{210}, 0) \cap & \\
\qquad (-\infty, -(a\eta_{120}\eta_{11} + c)/b\eta_{220}], & \\
\quad \eta_{12} \in [0, \eta_{120}] \cap (-\infty, -c/a\eta_{11}] \cap & \\
\qquad [-(b\eta_{220}\eta_{21} + c)/a\eta_{11}, \infty)\}, & \text{if } \eta_{120}\eta_{220} > 0 \text{ and } a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} \geq c.
\end{cases} \tag{5.53}
$$

□

Proof 5.7: *Let*

$$
\begin{aligned}
S_{01} &= \{(0, \eta_{12}, \eta_{21}, \eta_{22}) \in S\}, \\
S_{10} &= \{(\eta_{11}, \eta_{12}, 0, \eta_{22}) \in S\}, \\
S_{11} &= \{(\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in S : \eta_{11}\eta_{21} > 0\},
\end{aligned}
$$

so $S = S_{01} \cup S_{10} \cup S_{11}$.

First, suppose $(\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in S_{01}$*: then* $\eta_{11} = 0$ *and* $b\eta_{21}\eta_{22} + c = 0$*, which means that* $\eta_{21} < 0$*,* $\eta_{22} > 0$ *(so* $S_{01} = \emptyset$ *if* $\eta_{210}\eta_{220} = 0$*). Thus* $\eta_{22} = -c/b\eta_{21}$*. But* $\eta_{22} \in [0, \eta_{220}]$*, so* $\eta_{220} \geq -c/b\eta_{21}$ *which is equivalent to* $\eta_{21} \leq -c/b\eta_{220}$*. This in turn implies* $-\eta_{210} \leq -c/b\eta_{220}$ *or* $b\eta_{210}\eta_{220} \geq c$*. Clearly,* $S_{01} = \emptyset$ *if this last condition is not fulfilled, and* S_{01} *has the alternative form given if it is. The result given for* S_{10} *follows in a similar fashion.*

Now, let $(\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in S_{11}$*, and suppose* $\eta_{120}\eta_{220} > 0$*. As* $\eta_{22} = -(a\eta_{11}\eta_{12} + c)/b\eta_{21} \in [0, \eta_{220}]$*,* $a\eta_{11}\eta_{12} + c \geq 0$ *and so* $\eta_{12} \leq -c/a\eta_{11}$ *Also*

$$
\eta_{220} \geq -(a\eta_{11}\eta_{12} + c)/b\eta_{21} \quad \Leftrightarrow \quad \eta_{12} \geq -(b\eta_{220}\eta_{21} + c)/a\eta_{11}
$$

$$\Rightarrow \quad \eta_{120} \geq -(b\eta_{220}\eta_{21} + c)/a\eta_{11}$$

$$\Leftrightarrow \quad \eta_{21} \leq -(a\eta_{120}\eta_{11} + c)/b\eta_{220}$$

$$\Rightarrow \quad -\eta_{210} \leq -(a\eta_{120}\eta_{11} + c)/b\eta_{220}$$

$$\Leftrightarrow \quad \eta_{11} \leq (b\eta_{210}\eta_{220} - c)/a\eta_{120}$$

$$\Rightarrow \quad -\eta_{110} \leq (b\eta_{210}\eta_{220} - c)/a\eta_{120}$$

$$\Leftrightarrow \quad a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} \geq c.$$

Reading between the lines of this chain of implications and equivalences enables the deduction of the first and last forms for S_{11}. The other two forms follow by similar arguments. ■

Lemma 5.8:

$$T = \begin{cases} \emptyset, & \text{if } c > a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220}; \\ \{(\eta_{11},\eta_{21}) : \\ \quad \eta_{11} \in [-\eta_{110}, (b\eta_{210}\eta_{220} - c)/a\eta_{120}], \\ \quad \eta_{21} \in [-\eta_{210}, -(a\eta_{120}\eta_{11} + c)/b\eta_{220}]\}, & \text{if } a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} \geq c > \max\{a\eta_{110}\eta_{120}, b\eta_{210}\eta_{220}\}; \\ \{(\eta_{11},\eta_{21}) : \\ \quad \eta_{11} \in [-\eta_{110}, 0], \\ \quad \eta_{21} \in [-\eta_{210}, -(a\eta_{120}\eta_{11} + c)/b\eta_{220}]\}, & \text{if } b\eta_{210}\eta_{220} \geq c > a\eta_{110}\eta_{120}; \\ ([-\eta_{110}, -c/a\eta_{120}] \times [-\eta_{210}, 0]) \cup \\ \quad \{(\eta_{11},\eta_{21}) : \\ \quad\quad \eta_{11} \in (-c/a\eta_{120}, (b\eta_{210}\eta_{220} - c)/a\eta_{120}], \\ \quad\quad \eta_{21} \in [-\eta_{210}, -(a\eta_{120}\eta_{11} + c)/b\eta_{220}]\}, & \text{if } a\eta_{110}\eta_{120} \geq c > b\eta_{210}\eta_{220}; \\ ([-\eta_{110}, -c/a\eta_{120}] \times [-\eta_{210}, 0]) \cup \\ \quad \{(\eta_{11},\eta_{21}) : \\ \quad\quad \eta_{11} \in (-c/a\eta_{120}, 0], \\ \quad\quad \eta_{21} \in [-\eta_{210}, -(a\eta_{120}\eta_{11} + c)/b\eta_{220}]\}, & \text{if } \min\{a\eta_{110}\eta_{120}, b\eta_{210}\eta_{220}\} \geq c; \end{cases} \tag{5.54}$$

□

Proof 5.8: *Follows from the previous Lemma.* ■

The next thing to do is to find $\max\{\min\{n_{01}\xi_1 + n_{02}\eta_{11}, n_{01}\xi_2 + n_{02}\eta_{21}\} : (\eta_{11},\eta_{21}) \in T\}$.

Lemma 5.9: Let $B := \max\{\min\{n_{01}\xi_1 + n_{02}\eta_{11}, n_{01}\xi_2 + n_{02}\eta_{21}\} : (\eta_{11},\eta_{21}) \in T\}$.

If $c > a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220}$ then $B = -\infty$;

if $a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} \geq c > \max\{a\eta_{110}\eta_{120}, b\eta_{210}\eta_{220}\}$, then

$$B = \begin{cases} n_{01}\xi_1 - n_{02}\eta_{110}, & n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq -\frac{(a\eta_{120}+b\eta_{220})\eta_{110}-c}{b\eta_{220}}; \\ \frac{n_{01}(a\eta_{120}\xi_1+b\eta_{220}\xi_2)-n_{02}c}{a\eta_{120}+b\eta_{220}}, & -\frac{(a\eta_{120}+b\eta_{220})\eta_{110}-c}{b\eta_{220}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq \\ & \quad \frac{(a\eta_{120}+b\eta_{220})\eta_{210}-c}{a\eta_{120}}; \\ n_{01}\xi_1 + n_{02}\frac{b\eta_{210}\eta_{220}-c}{a\eta_{120}}, & \frac{(a\eta_{120}+b\eta_{220})\eta_{210}-c}{a\eta_{120}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1); \end{cases} \tag{5.55}$$

if $b\eta_{210}\eta_{220} \geq c > a\eta_{110}\eta_{120}$ then

$$B = \begin{cases} n_{01}\xi_1 - n_{02}\eta_{110}, & n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq -\frac{(a\eta_{120}+b\eta_{220})\eta_{110}-c}{b\eta_{220}}; \\ \frac{n_{01}(a\eta_{120}\xi_1+b\eta_{220}\xi_2)-n_{02}c}{a\eta_{120}+b\eta_{220}}, & -\frac{(a\eta_{120}+b\eta_{220})\eta_{110}-c}{b\eta_{220}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq \frac{c}{b\eta_{220}}; \\ n_{01}\xi_1, & \frac{c}{b\eta_{220}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1); \end{cases} \tag{5.56}$$

if $a\eta_{110}\eta_{120} \geq c > b\eta_{210}\eta_{220}$ then

$$B = \begin{cases} n_{01}\xi_1 - n_{02}\frac{c}{a\eta_{120}}, & n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq -\frac{c}{a\eta_{120}}; \\ \frac{n_{01}(a\eta_{120}\xi_1+b\eta_{220}\xi_2)-n_{02}c}{a\eta_{120}+b\eta_{220}}, & -\frac{c}{a\eta_{120}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq \frac{(a\eta_{120}+b\eta_{220})\eta_{210}-c}{a\eta_{120}}; \\ n_{01}\xi_1 + n_{02}\frac{b\eta_{210}\eta_{220}-c}{a\eta_{120}}, & \frac{(a\eta_{120}+b\eta_{220})\eta_{210}-c}{a\eta_{120}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1); \end{cases} \tag{5.57}$$

if $\min\{a\eta_{110}\eta_{120}, b\eta_{210}\eta_{220}\} \geq c$ then

$$B = \begin{cases} n_{01}\xi_1 - n_{02}\frac{c}{a\eta_{120}}, & n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq -\frac{c}{a\eta_{120}}; \\ \frac{n_{01}(a\eta_{120}\xi_1+b\eta_{220}\xi_2)-n_{02}c}{a\eta_{120}+b\eta_{220}}, & -\frac{c}{a\eta_{120}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1) \leq \frac{c}{b\eta_{220}}; \\ n_{01}\xi_1, & \frac{c}{b\eta_{220}} \leq n_{02}^{-1}n_{01}(\xi_2 - \xi_1). \end{cases} \tag{5.58}$$

□

Proof 5.9: *Follows by some tedious algebra after noting that* $B = \max\{n_{01}\xi_1 + \max\{n_{02}\eta_1 : n_{02}\eta_1 \leq n_{02}\eta_2 + n_{01}(\xi_2 - \xi_1), (\eta_1, \eta_2) \in T\}, n_{01}\xi_2 + \max\{n_{02}\eta_2 : n_{02}\eta_2 \leq n_{02}\eta_1 - n_{01}(\xi_2 - \xi_1), (\eta_1, \eta_2) \in T\}\}$, *so* $B = \max\{n_{01}\xi_1 + n_{02}\max\{\eta_1 : \eta_1 \leq \eta_2 + n_{02}^{-1}n_{01}(\xi_2 - \xi_1)\}, n_{01}\xi_2 + n_{02}\max\{\eta_2 : \eta_2 \leq \eta_1 - n_{02}^{-1}n_{01}(\xi_2 - \xi_1)\}\}$ *as* $n_{02} > 0$. ■

Corollary 5.10: Let B be defined as in Lemma 5.9. Then, if $a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} \geq c$,

$$\begin{aligned} B \leq \max\Bigg\{ & \min\left\{ n_{01}\xi_1 + n_{02}\frac{b\eta_{210}\eta_{220} - c}{a\eta_{120}}, n_{01}\xi_2 - n_{02}\eta_{210} \right\}, \\ & \min\left\{ n_{01}\xi_1 - n_{02}\eta_{110}, n_{01}\xi_2 + n_{02}\frac{a\eta_{110}\eta_{120} - c}{b\eta_{220}} \right\} \Bigg\}. \end{aligned} \tag{5.59}$$

□

Proof 5.10: *Although this Corollary can be derived from the previous Lemma, it follows immediately from the fact that, if* $a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} \geq c$, *then* $\left(\frac{b\eta_{210}\eta_{220}-c}{a\eta_{120}}, -\eta_{210}\right), \left(-\eta_{110}, \frac{a\eta_{110}\eta_{120}-c}{b\eta_{220}},\right) \in T$. ■

Now some necessary conditions on the value of ζ_1 such that $\mathcal{E}_2 = \mathcal{E}_1$ for given ξ_1, ξ_2 and ζ_2 (where $\zeta_2 > 0$) can be derived. There are three cases where this can be done:

1. the intersections of the hyperplanes $\mathbb{H}_1^{\pm}$ with the hyperplane $\mathbb{H}_2^{+}$ both fall outside of the interior of $\mathcal{E}_0$; this is the case $n_{02}^{-1}(\zeta_2 - n_{01}\xi_i) \geq \sqrt{1 - \xi_i^2}$, $i = 1, 2$ or $\zeta_2 > \max_{i=1,2}\left\{n_{01}\xi_i + n_{02}\sqrt{1 - \xi_i^2}\right\}$;

2. one of the intersections $\mathbb{H}_1^{\pm} \cap \mathbb{H}_2^{+}$ falls in the interior of $\mathcal{E}_0$, but the hyperplane $\mathbb{H}_2^{+}$ does not cross the "equatorial" hyperplane between $\mathbb{H}_1^{+}$ and $\mathbb{H}_1^{-}$; that is, $\min_{i=1,2}\left\{n_{02}^{-1}(\zeta_2 - n_{01}\xi_i)\right\} \geq 0$, so $\max_{i=1,2}\left\{n_{01}\xi_i + n_{02}\sqrt{1-\xi_i^2}\right\} > \zeta_2 \geq \min_{i=1,2}\left\{n_{01}\xi_i + n_{02}\sqrt{1-\xi_i^2}\right\}$ and $\zeta_2 \geq \max_{i=1,2}\{n_{01}\xi_i\}$;

3. both of the intersections $\mathbb{H}_1^{\pm} \cap \mathbb{H}_2^{+}$ fall in the interior of $\mathcal{E}_0$, but the hyperplane $\mathbb{H}_2^{+}$ again does not cross the "equatorial" hyperplane between $\mathbb{H}_1^{+}$ and $\mathbb{H}_1^{-}$; so $\min_{i=1,2}\left\{n_{01}\xi_i + n_{02}\sqrt{1-\xi_i^2}\right\} > \zeta_2 \geq \max_{i=1,2}\{n_{01}\xi_i\}$;

There will be theorems giving a range of values of ζ_1 such that $\mathcal{E}_2 = \mathcal{E}_1$ for these cases.

Theorem 5.11: Suppose

$$\zeta_2 > \max_{i=1,2}\left\{n_{01}\xi_i + n_{02}\sqrt{1-\xi_i^2}\right\}.$$

Then, if

$$\zeta_1 \leq B, \tag{5.60}$$

$\mathcal{E}_2 = \mathcal{E}_1$, where B is given by one of the equations (5.56) to (5.58) with $\eta_{110} = \eta_{120} = \sqrt{1-\xi_1^2}$, $\eta_{210} = \eta_{220} = \sqrt{1-\xi_2^2}$, $a = \xi_2 - \xi_{01}$, $b = \xi_{01} - \xi_1$, $c = \left[\left(1 + \xi_1\xi_2 - (\xi_1 + \xi_2)\xi_{01}\right)(\xi_2 - \xi_1)\right]/(p-1)$ and ξ_{01} given by equation (5.21), as follows:

if $p = 3$: B is given by equation

(5.56) if $\xi_1 \in (-1, 0)$ and $\xi_2 \in (\xi_1, -\xi_1)$;

(5.57) if $\xi_1 \in (-1, 0)$ and $\xi_2 \in (-\xi_1, 1)$, or if $\xi_1 \in [0, 1)$;

(5.58) if $\xi_1 \in (-1, 0)$ and $\xi_2 = -\xi_1$.

if $p > 3$: B is given by equation

(5.56) if $\xi_1 \in \left(-1, -\sqrt{\frac{p-3}{2(p-2)}}\right)$ and $\xi_2 \in \left(-\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}, \sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}\right)$;

(5.57) if $\xi_1 \in \left[-\sqrt{\frac{p-3}{2(p-2)}}, 1\right)$ and $\xi_2 \in \left(\sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}, 1\right)$;

(5.58) if $\xi_1 \in \left(-1, -\sqrt{\frac{p-3}{2(p-2)}}\right)$ and $\xi_2 \in \left(\xi_1, -\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}\right] \cup \left[\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}, \frac{1}{p}\sqrt{\frac{2(p-2)}{p-3}}\right]$, or if $\xi_1 \in \left[-\sqrt{\frac{p-3}{2(p-2)}}, 1\right)$ and $\xi_2 \in \left(\xi_1, \sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}\right]$.

□

Proof 5.11: *The idea is to find a set of support points for the distribution of Theorem 5.5. The position of the hyperplanes $\mathbb{H}_1^{\pm}$ and $\mathbb{H}_2^{+}$ means that possible values for the η's are such that*

$(\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in [-\eta_{110}, 0] \times [0, \eta_{120}] \times [-\eta_{210}, 0] \times [0, \eta_{220}]$, *so long as equation (5.23) holds (with* ξ_{01} *given by equation (5.21)), where* $\eta_{i10} = \eta_{i20} = \sqrt{1-\xi_i^2}$, $i = 1, 2$, *and* ζ_1 *satisfies* $\zeta_1 \leq \min_{i=1,2}\{n_{01}\xi_i + n_{02}\eta_{i1}\}$. *But, if* $(\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in [-\eta_{110}, 0] \times [0, \eta_{120}] \times [-\eta_{210}, 0] \times [0, \eta_{220}]$, *and equation (5.23) holds,* $(\eta_{11}, \eta_{21}) \in T$, *with* $a = \xi_2 - \xi_{01}$, $b = \xi_0 - \xi_{01}$, *and* $c = [(1 + \xi_1\xi_2 - (\xi_1 + \xi_2)\xi_{01})(\xi_2 - \xi_1)]/(p-1)$. *Then, the relation*

$$\zeta_1 \leq \max\left\{\min_{i=1,2}\{n_{01}\xi_i + n_{02}\eta_{i1}\} : (\eta_{11}, \eta_{21}) \in T\right\} \tag{5.61}$$

is sufficient to ensure that $\mathcal{E}_2 = \mathcal{E}_1$.

Now, $a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} = (p-1)c > c$, *as* $c > 0$, *so, to find which form from Lemma 5.9 needs to be used it is necessary to determine the ordering of the quantities* $a\eta_{110}\eta_{120}$, $b\eta_{210}\eta_{220}$ *and* c.

For the moment, assume $p > 3$ *and let*

$$\begin{aligned} f(\xi_1, \xi_2) &= c - a\eta_{110}\eta_{120} \\ &= \frac{(\xi_2 - \xi_1)(1 + \xi_1\xi_2 - (\xi_1 + \xi_2)\xi_{01})}{p-1} - (\xi_2 - \xi_{01})(1 - \xi_1^2). \end{aligned}$$

As ξ_{01} *is a continuous function (provided it is redefined to be* 0 *at its removable singularity* $\xi_2 = \xi_1$*) of* ξ_1 *and* ξ_2, f *is a continuous function of* ξ_1 *and* ξ_2. *Considered as a function of* ξ_2, f *has zeros at*

$$\xi_2 = \pm\sqrt{\frac{2(p-2)\xi_1^2 - (p-3)}{p-1}},\ \xi_2 = \xi_1$$

and these are all real if $\xi_1^2 \geq (p-3)/2(p-2)$.

Also,

$$\xi_0 = \frac{\pm 1 + p\xi_1}{p+1} \text{ when } \xi_2 = \pm 1$$

so

$$f(\xi_1, \pm 1) = \frac{\pm p^2(1-\xi_1^2)(1 \mp \xi_1)}{p+1}$$

and $f(\xi_1, 1) < 0$, $f(\xi_1, -1) > 0$, *for* $\xi_1 \in (-1, 1)$.

If $\xi_2 = \xi_1 + \epsilon$, $f(\xi_1, \xi_2) = -\frac{p-2}{p-1}(1-\xi_1^2)\epsilon + \mathrm{O}(\epsilon^2)$, *so* $f(\xi_1, \xi_2)$, *considered as a function of* ξ_2, *is decreasing as it passes through* $\xi_2 = \xi_1$ *for all* $\xi_1 \in (-1, 1)$.

Also,

$$\begin{aligned} \left.\frac{\partial}{\partial\xi_2} f(\xi_1, \xi_2)\right|_{\xi_2 = \sqrt{\frac{2(p-2)\xi_1^2 - (p-3)}{p-1}}} &= \frac{4(p-1)\sqrt{2(p-2)\xi_1^2 - (p-3)}}{(p+1)(p^2 - 3p + 4)} \\ &\quad \times \left[\sqrt{p-1}\xi_1 - \sqrt{2(p-2)\xi_1^2 - (p-3)}\right] \end{aligned}$$

which (when real) has the same sign as ξ_1.

This information about f *at* $(\xi_1, \pm 1)$, (ξ_1, ξ_1) *and* $\left(\xi_1, \sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}\right)$, *the continuity of* f *and the fact that* $\xi_1 \leq -\sqrt{\frac{p-3}{2(p-2)}} \Rightarrow \xi_1 < -\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}$, *and* $\xi_1 \geq \sqrt{\frac{p-3}{2(p-2)}} \Rightarrow \xi_1 > \sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}$ *allows the behaviour of* f *to be deduced in outline:*

if $\xi_1 < -\sqrt{\frac{p-3}{2(p-2)}}$, f *is positive for* $\xi_2 \in \left(-\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}, \sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}\right)$ *and non-positive for* $\xi_2 \in \left(\xi_1, -\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}\right] \cup \left[\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}, 1\right)$;

if $\xi_1 \geq -\sqrt{\frac{p-3}{2(p-2)}}$, f *is non-positive for* $\xi_2 \in (\xi_1, 1)$.

Now let $g(\xi_1, \xi_2) = c - b\eta_{210}\eta_{220}$. g *is continuous, and, considered as a function of* ξ_2, *has zeros*

$$\xi_2 = \pm\sqrt{\frac{(p-1)\xi_1^2 + p - 3}{2(p-2)}}, \ \xi_2 = \xi_1,$$

which are always real.

Also,

$$g(\xi_1, \pm 1) = \frac{p(\pm 1 - \xi_1)(1 - \xi_1^2)}{p^2 - 1},$$

$$g(\xi_1, \xi_1 + \epsilon) = -\frac{(p-3)(1-\xi_1^2)}{2(p-1)}\epsilon + \mathrm{O}(\epsilon^2).$$

These relations, together with the facts that g *is continuous and* $|\xi_1| \leq \sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}$ *are enough to deduce that* $g(\xi_1, \xi_2)$ *is non-positive for* $\xi_2 \in \left(\xi_1, \sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}\right]$ *and positive for* $\xi_2 \in \left(\sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}, 1\right)$.

Hence, when

1. $\xi_1 < -\sqrt{\frac{p-3}{2(p-2)}}$,

   (a) $b\eta_{210}\eta_{220} \geq c > a\eta_{110}\eta_{120}$ *for* $\xi_2 \in \left(-\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}, \sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}\right)$;

   (b) $\min\{a\eta_{110}\eta_{120}, b\eta_{210}\eta_{220}\} \geq c$ *for* $\xi_2 \in \left(\xi_1, -\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}\right] \cup \left[\sqrt{\frac{2(p-2)\xi_1^2-(p-3)}{p-1}}, \frac{1}{p}\sqrt{\frac{2(p-2)}{p-3}}\right]$.

   It will be noted that the condition $\xi_1\xi_2 > -1/p$ *implies that* $\xi_2 < \frac{1}{p}\sqrt{\frac{2(p-2)}{p-3}} < \sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}$;

2. $\xi_1 \geq -\sqrt{\frac{p-3}{2(p-2)}}$,

(a) $a\eta_{110}\eta_{120} \geq c > b\eta_{210}\eta_{220}$ *for* $\xi_2 \in \left(\sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}, 1\right)$;

(b) $\min\{a\eta_{110}\eta_{120}, b\eta_{210}\eta_{220}\} \geq c$ *for* $\xi_2 \in \left(\xi_1, \sqrt{\frac{(p-1)\xi_1^2+p-3}{2(p-2)}}\right]$.

Now the Theorem for $p > 3$ *follows directly from Lemma 5.9.*

Suppose now $p = 3$. *Then* $g = -f$ *and* f *considered as a function of* ξ_2 *has the zeros* $\xi_2 = \pm\xi_1$. *Also,* $f(\xi_1, 1) = -3(1-\xi_1)(1-\xi_1^2)/8 < 0$ *for* $\xi_1 \in (-1, 1)$,

$$\frac{\partial}{\partial\xi_2} f(\xi_1,\xi_2)\Big|_{\xi_2=\xi_1} = 0, \quad \frac{\partial^2}{\partial\xi_2^2} f(\xi_1,\xi_2)\Big|_{\xi_2=\xi_1} = -2\xi_1, \quad \frac{\partial^3}{\partial\xi_2^3} f(\xi_1,\xi_2)\Big|_{\xi_2=\xi_1} = -3$$

and

$$\frac{\partial}{\partial\xi_2} f(\xi_1,\xi_2)\Big|_{\xi_2=-\xi_1} = -2\xi_1^2,$$

and these facts, together with the continuity of f, *are sufficient to deduce that*

if $\xi_1 \in (-1, 0)$, f *is positive for* $\xi_2 \in (\xi_1, -\xi_1)$, *and non-positive for* $\xi_2 \in [-\xi_1, 1)$;

if $\xi_1 \in [0, 1)$, f *non-positive for* $\xi_2 \in [\xi_1, 1)$.

Together with $g = -f$, *the above is sufficient to allow the deduction of the rest of the Theorem.*

■

Theorem 5.12: Suppose ζ_2 obeys one of the sets of conditions given in the cases 1, 2 or 3 on page 146. Then, if

$$\begin{aligned}\zeta_1 \quad \leq \quad & \max\left\{\min\left\{n_{01}\xi_1 + n_{02}\frac{b\eta_{210}\eta_{220}-c}{a\eta_{120}}, n_{01}\xi_2 - n_{02}\eta_{210}\right\},\right. \\ & \left.\min\left\{n_{01}\xi_1 - n_{02}\eta_{110}, n_{01}\xi_2 + n_{02}\frac{a\eta_{110}\eta_{120}-c}{b\eta_{220}}\right\}\right\}, \qquad (5.62)\end{aligned}$$

where $a = \xi_2 - \xi_{01}$, $b = \xi_{01} - \xi_1$, $c = (\xi_2 - \xi_1)[1 + \xi_1\xi_2 - (\xi_1 + \xi_2)\xi_{01}]/(p-1)$, $\eta_{110} = \eta_{120} = \sqrt{1-\xi_1^2}$ and, in

case 1: $\eta_{210} = \eta_{220} = \sqrt{1-\xi_2^2}$;

case 2: let j be such that $n_{01}\xi_j + n_{02}\sqrt{1-\xi_j^2} = \max_{i=1,2} n_{01}\xi_i + n_{02}\sqrt{1-\xi_i^2}$, $\ell \in \{1,2\} - \{j\}$.
Then $\eta_{2j0} = n_{02}^{-1}(\zeta_2 - n_{01}\xi_j)$, $\eta_{2\ell 0} = \sqrt{1-\xi_2^2}$;

case 3: $\eta_{210} = n_{02}^{-1}(\zeta_2 - n_{01}\xi_1)$ and $\eta_{220} = n_{02}^{-1}(\zeta_2 - n_{01}\xi_2)$.

□

Proof 5.12: *In all cases,* $\sqrt{1-\xi_i^2} \geq \eta_{2i0}$, *so* $a\eta_{110}\eta_{120} + b\eta_{210}\eta_{220} \geq a(1-\xi_1^2) + b(1-\xi_2^2) = (p-1)c > c$. *But then Corollary 5.10 applies and the distribution given by putting* $\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}$ *in the quantities of Theorem 5.5 will have support points contained in* $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ *if* $(\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22}) \in S$. *Thus,* $\mathcal{E}_2 = \mathcal{E}_0$. ■

5.2.2 $\mathcal{E}_1 = \mathcal{E}(0, I)$

Now $\xi_1\xi_2 \leq -1/p$ and $\xi_{01} = 0$. However, the following Theorem holds by the methods of Theorem 5.5:

Theorem 5.13: If $\xi_1\xi_2 \leq -1/p$, suppose $\xi_2' \in [-1/p\xi_1, \xi_2]$. Let $\xi_1' = -1/p\xi_2'$ and $x_{ijkl} = \xi_i' e_1 + \eta_{ij} e_2 + (-1)^k \sqrt{1 - \xi_i'^2 - \eta_{ij}^2}\, e_{l+2}$, $i, j, k = 1, 2$, $l = 1, \ldots, (p-2)$, for some η_{ij} such that $\eta_{ij} \in [-\sqrt{1-\xi_i'^2}, \sqrt{1-\xi_i'^2}]$ $\forall i, j = 1, 2$ (so that $x_{ijkl} \in \mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$ $\forall i, j, k = 1, 2, l = 1, \ldots, p-1$) be the support points of a distribution w which takes on the nonnegative values w_{ijkl} at these support points.

Then, if

$$\eta_{11} \leq \eta_{12} \quad \text{and} \quad \eta_{21} \leq \eta_{22}, \tag{5.63}$$

the η's also satisfy

$$-\xi_2'\eta_{11}\eta_{12} + \xi_1'\eta_{21}\eta_{22} = \frac{1 + \xi_1'\xi_2'}{p-1}(\xi_2' - \xi_1') \tag{5.64}$$

and the weights w_{ijkl} are given by

$$w_{ij11} = \frac{\xi_2'\eta_{12}}{2(p-2)(\xi_2' - \xi_1')(\eta_{12} - \eta_{11})} \tag{5.65}$$

$$w_{ij12} = \frac{-\xi_2'\eta_{11}}{2(p-2)(\xi_2' - \xi_1')(\eta_{12} - \eta_{11})} \tag{5.66}$$

$$w_{ij21} = \frac{-\xi_1'\eta_{22}}{2(p-2)(\xi_2' - \xi_1')(\eta_{22} - \eta_{21})} \tag{5.67}$$

$$w_{ij22} = \frac{\xi_1'\eta_{21}}{2(p-2)(\xi_2' - \xi_1')(\eta_{22} - \eta_{21})}, \quad \forall i, j = 1, 2, \tag{5.68}$$

the distribution w is optimal for $\mathcal{E}_0 \cap \bar{\Pi}(e_1, \xi_1, \xi_2)$. □

The following Corollary can be deduced from the preceding Theorem.

Corollary 5.14: Suppose ζ_1 obeys

$$\zeta_1 \leq \max\{B_1, B_2\}, \tag{5.69}$$

where

$$B_1 = \max\{\min\{n_{01}\xi_1'(\xi_2') + n_{02}\eta_{11}(\xi_2'), n_{01}\xi_2' - n_{02}\eta_{210}(\xi_2')\}: \eta_{210}(\xi_2'), \eta_{220}(\xi_2') > 0, \eta_{11}(\xi_2') \in [-\eta_{110}(\xi_2'), 0)\} \tag{5.70}$$

for

$$\eta_{11}(\xi_2') = \frac{b(\xi_2')\eta_{210}(\xi_2')\eta_{220}(\xi_2') - c(\xi_2')}{a(\xi_2')\eta_{120}(\xi_2')} \tag{5.71}$$

and

$$B_2 = \max\{\min\{n_{01}\xi_1'(\xi_2') - n_{02}\eta_{110}(\xi_2'), n_{01}\xi_2' + n_{02}\eta_{21}(\xi_2')\}: \eta_{210}(\xi_2'), \eta_{220}(\xi_2') > 0, \eta_{21}(\xi_2') \in [-\eta_{210}(\xi_2'), 0)\} \quad (5.72)$$

for

$$\eta_{11}(\xi_2') = \frac{b(\xi_2')\eta_{210}(\xi_2')\eta_{220}(\xi_2') - c(\xi_2')}{a(\xi_2')\eta_{120}(\xi_2')}, \quad (5.73)$$

and

$$\eta_{110} = \sqrt{1 - \xi_1'(\xi_2')^2}, \quad (5.74)$$

$$\eta_{120} = \min\left\{\sqrt{1 - \xi_1'(\xi_2')^2}, n_{02}^{-1}(\zeta_2 - n_{01}\xi_1'(\xi_2'))\right\}, \quad (5.75)$$

$$\eta_{210} = \sqrt{1 - \xi_2'^2}, \quad (5.76)$$

$$\eta_{220} = \min\left\{\sqrt{1 - \xi_2'^2}, n_{02}^{-1}(\zeta_2 - n_{01}\xi_2')\right\}, \quad (5.77)$$

$$\xi_1'(\xi_2') = \frac{-1}{p\xi_2'}, \quad (5.78)$$

$$a = \xi_2', \quad (5.79)$$

$$b = -\xi_1'(\xi_2'), \quad (5.80)$$

$$c = \frac{(\xi_2' - \xi_1'(\xi_2'))(1 + \xi_2'\xi_1'(\xi_2'))}{p - 1} = \frac{p\xi_2'^2 + 1}{p^2\xi_2'}, \quad (5.81)$$

then $\mathcal{E}_2 = \mathcal{E}_0$. □

Proof 5.14:

With the values given by the statement of the Corollary, the distribution with support points $x_{11kl} = \xi_1'(\xi_2')e_1 + \eta_{11}(\xi_2')e_2 + (-1)^k\sqrt{1 - \xi_1'(\xi_2')^2 - \eta_{11}(\xi_2')^2}e_{l+2}$, $x_{12kl} = \xi_1'(\xi_2')e_1 - \eta_{120}(\xi_2')e_2 + (-1)^k\sqrt{1 - \xi_1'(\xi_2')^2 - \eta_{120}(\xi_2')^2}e_{l+2}$, $x_{21kl} = \xi_2'e_1 + \eta_{210}(\xi_2')e_2 + (-1)^k\sqrt{1 - \xi_2'^2 - \eta_{210}(\xi_2')^2}e_{l+2}$, $x_{22kl} = \xi_2'e_1 + \eta_{220}(\xi_2')e_2 + (-1)^k\sqrt{1 - \xi_2'^2 - \eta_{220}(\xi_2')^2}e_{l+2}$, *and weights*

$$w_{ij11} = \frac{\xi_2'\eta_{120}(\xi_2')}{2(p-2)(\xi_2' - \xi_1'(\xi_2'))(\eta_{120}(\xi_2') - \eta_{11}(\xi_2'))} \quad (5.82)$$

$$w_{ij12} = \frac{-\xi_2'\eta_{11}(\xi_2')}{2(p-2)(\xi_2' - \xi_1'(\xi_2'))(\eta_{120}(\xi_2') - \eta_{11}(\xi_2'))} \quad (5.83)$$

$$w_{ij21} = \frac{-\xi_1'(\xi_2')\eta_{220}(\xi_2')}{2(p-2)(\xi_2' - \xi_1'(\xi_2'))(\eta_{220}(\xi_2') + \eta_{210}(\xi_2'))} \quad (5.84)$$

$$w_{ij22} = \frac{-\xi_1'(\xi_2')\eta_{210}(\xi_2')}{2(p-2)(\xi_2' - \xi_1'(\xi_2'))(\eta_{220}(\xi_2') + \eta_{210}(\xi_2'))}, \ \forall i, j = 1, 2, \quad (5.85)$$

is optimal for both $\mathcal{E}_i$ and $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ provided $\eta_{210}(\xi_2'), \eta_{220}(\xi_2') > 0$, $\eta_{11}(\xi_2') \in [-\eta_{110}(\xi_2'), 0)$ and $\zeta_1 \leq {n_0}^{\mathrm{T}} x$ for all support points x of this distribution. But the last condition is guaranteed by $\zeta_1 \leq B_1$, so $\zeta_1 \leq B_1$ implies $\mathcal{E}_2 = \mathcal{E}_0$.

Similarly, $\zeta_1 \leq B_2$ implies $\mathcal{E}_2 = \mathcal{E}_0$. ■

5.3 Contact Points

In this section, it will be assumed that $\mathcal{E}_2 \neq \mathcal{E}_0$.

Some definitions are necessary here:

Definition 5.2: *For a set S let S° denote its topological interior and let $\partial S = S - S^\circ$ be its boundary.* □

Here a condition on the centre of $\mathcal{E}_2$ will be given.

Lemma 5.15: If (ξ_0, η_0) is the centre of $\mathcal{E}_2$, then

$$(\xi_0, \eta_0) \in (\mathcal{E}_0 \cap \Pi_1 \cap_2)^\circ. \tag{5.86}$$

□

Proof 5.15: *Suppose $(\xi_0, \eta_0) \notin (\mathcal{E}_0 \cap \Pi_1 \cap_2)^\circ$. Then there exists $n \in \mathbb{R}^2$, $k \in \mathbb{R}$ such that the half-space $H = \{z \in \mathbb{R}^p : [n^{\mathrm{T}}, 0]z \leq k\} \supset \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ and $(\xi_0, \eta_0) \in \overline{\mathbb{R}^p - H}$. But then the minimum-volume ellipsoid $\mathcal{E}_2'$ around the set $H \cap \mathcal{E}_2$ is of smaller volume than $\mathcal{E}_2$ and contains $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$. This contradiction proves the Lemma.* ■

Now it is necessary to give some definitions: first a projection is defined, and then the contact points between $\mathcal{E}_0$ and another ellipsoid $\mathcal{E}$ will be classified.

Definition 5.3: *Let π be the projection $\pi : (\xi, \eta, x) \in \mathbb{R}^p \mapsto (\xi, \eta) \in \mathbb{R}^2$.* □

Definition 5.4: *$\partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2) = (\partial\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2) \cup (\mathcal{E}_0 \cap \mathbb{H}_1^+ \cap \Pi_2) \cup (\mathcal{E}_0 \cap \mathbb{H}_1^- \cap \Pi_2) \cup (\mathcal{E}_0 \cap \Pi_1 \cap \mathbb{H}_2^+) \cup (\mathcal{E}_0 \cap \Pi_1 \cap \mathbb{H}_2^-)$ and $\mathcal{E}_0 \cap \Pi_i \cap \mathbb{H}_j^\pm \supset \mathcal{E}_0 \cap \mathbb{H}_i^{\pm'} \cap \mathbb{H}_j^\pm$ for $i, j \in \{1, 2\}, j \neq i$.*

Let $R_2 = \bigcup\{\mathcal{E}_0 \cap \mathbb{H}_1^\pm \cap \mathbb{H}_2^{\pm'} : \mathbb{H}_1^\pm \cap \mathcal{E}_0^\circ \neq \emptyset, \mathbb{H}_2^{\pm'} \cap \mathcal{E}_0^\circ \neq \emptyset\}$. Let $R_1 = \bigcup\{\mathcal{E}_0 \cap \Pi_i \cap \mathbb{H}_j^\pm : (i, j) \in \{(1, 2), (2, 1)\}, \mathcal{E}_0^\circ \cap \mathbb{H}_j^\pm \neq \emptyset\} - R_2$. Let $R_0 = \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2) - R_1 - R_2$.

*Let $\mathcal{E}$ be an ellipsoid. If $(\xi, \eta, x) \in \partial\mathcal{E} \cap \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$ for some $x \in \mathbb{R}^{p-2}$, then (ξ, η) is a **contact point for** $\mathcal{E}$. If $(\xi, \eta, x) \in \partial\mathcal{E} \cap R_i$ for some $x \in \mathbb{R}^{p-2}$, then (ξ, η) is a **(contact) point of Type i for** $\mathcal{E}$. If (ξ, η) is a point of Type i for $\mathcal{E}$ and $\xi^2 + \eta^2 < 1$, (ξ, η) is a **(contact) point of Type ia for** $\mathcal{E}$. If (ξ, η) is a point of Type i for $\mathcal{E}$ and $\xi^2 + \eta^2 = 1$, (ξ, η) is a **(contact) point of Type ib for** $\mathcal{E}$.*

If "for $\mathcal{E}$" is omitted when the above definitions are used, "for $\mathcal{E}_2$" is to be understood. □

Thus, for example, if (ξ, η) is a point of Type 2, it corresponds to an equivalence class of points in $\mathbb{R}^p$ which belong to $\mathcal{E}_0 \cap \mathcal{E}_2$ and to one of the four sets $\mathbb{H}_1^{\pm} \cap \mathbb{H}_2^{\pm'}$, which consist of the intersection of *two* hyperplanes.

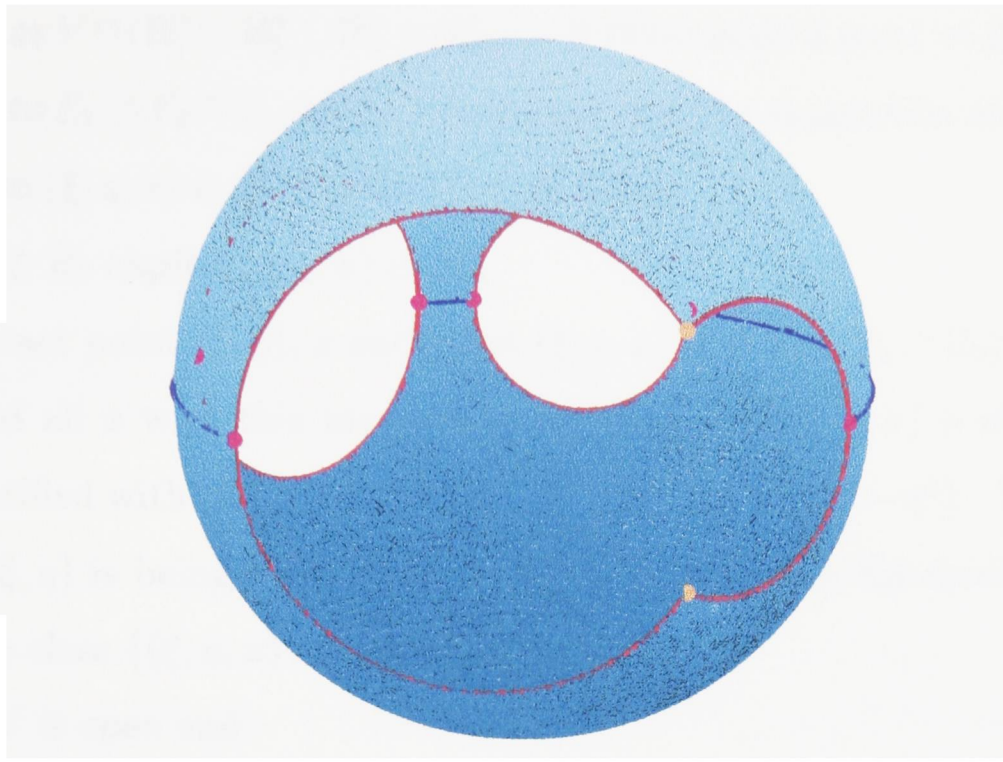


Figure 5.1: Possible contact points of Types 0, 1 and 2 in three dimensions. The possible points of Type 0a (which do not in fact occur) are cyan-coloured, those of Type 0b are blue, those of Type 1a are red, those of Type 1b are pink and the point of Type 2a is orange. Note that two points of the three-dimensional space in the diagram correspond to each point of Type ia, and one point of the diagram corresponds to each point of Type ib.

Some properties of points of Type 0, 1 and 2 will now be derived.

Theorem 5.16: If (ξ, η) is a point of Type 0, 1 or 2, then $(\xi, \eta, x) \in \partial\mathcal{E}_0$ for some $x \in \mathbb{R}^{p-2}$. Moreover, $(\xi, \eta, x) \in \partial\mathcal{E}_0$ for every $x \in \mathbb{R}^{p-2}$ such that $x^{\mathrm{T}}x = 1 - \xi^2 - \eta^2$. □

Proof 5.16: *The symmetry of $\mathcal{E}_0$, Π_1, Π_2 and $\mathcal{E}_2$ under rotations which leave e_1 and e_2 unaltered means that if $(\xi, \eta, x) \in \mathcal{E}_0 \cap \mathcal{E}_2$ belongs to i of the hyperplanes $\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-$, then $(\xi, \eta, x') \in \mathcal{E}_0 \cap \mathcal{E}_2$ belongs to the same i hyperplanes for every $x' \in \mathbb{R}^{p-2}$ such that $x'^{\mathrm{T}}x' = x^{\mathrm{T}}x$. Thus, the "Moreover" part of the Theorem follows from the main part.*

If (ξ, η) is a point of Type 0, then $\exists x \in \mathbb{R}^{p-2}$ such that $(\xi, \eta, x) \in \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2) - (\mathbb{H}_1^+ \cup \mathbb{H}_1^- \cup \mathbb{H}_2^+ \cup \mathbb{H}_2^-)$, so, obviously, $(\xi, \eta, x) \in \partial\mathcal{E}_0$.

Suppose now that (ξ, η) is a point of Type 1, and that there exists $x \in \mathbb{R}^{p-2}$ such that $(\xi, \eta, x) \in R_1 - \partial\mathcal{E}_0$. For the sake of definiteness, suppose $(\xi, \eta, x) \in \mathbb{H}_1^+$. Since $\partial\mathcal{E}_0 = \{w \in \mathbb{R}^p : w^{\mathrm{T}}w = 1\}$ and it can be assumed without loss of generality that $\mathbb{H}_1^+ = \{w \in \mathbb{R}^p : {e_1}^{\mathrm{T}}w = w_1 = \text{constant}\}$, there exists a neighbourhood V of (ξ, η, x) such that $V \cap (\mathbb{H}_1^- \cup \mathbb{H}_2^+ \cup \mathbb{H}_2^- \cup \partial\mathcal{E}_0) = \emptyset$.

Since $(\xi,\eta,x) \in \mathbb{H}_1^+$, $\xi = w_1$. *The point* $(w_1,\eta,x+\epsilon_x) \in \mathbb{H}_1^+ \cap V$ *if* ${\epsilon_x}^{\mathrm{T}}\epsilon_x$ *is sufficiently small, but* $(\xi,\eta)\bar{Q}^{-1}(\xi,\eta)^{\mathrm{T}} + \delta(x+\epsilon_x)^{\mathrm{T}}(x+\epsilon_x) = 1 + \delta(2{\epsilon_x}^{\mathrm{T}}x + {\epsilon_x}^{\mathrm{T}}\epsilon_x) \leq 1$ *for* ϵ_x *with sufficiently small, but nonzero, magnitude and arbitrary direction in the appropriate* $(p-2)$*-dimensional subset only if* $\delta = 0$, *which would contradict the positive definiteness of* Q. *Thus there exists a point of* $\mathbb{H}_1^+ \cap V - \mathcal{E}_2$, *and, as* $V \cap (\mathbb{H}_1^- \cup \mathbb{H}_2^+ \cup \mathbb{H}_2^- \cup \partial\mathcal{E}_0) = \emptyset$ *there exists a point of* $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2 - \mathcal{E}_2$, *which is in contradiction to* $\mathcal{E}_2 \supset \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$. *This means that the assumption that* $(\xi,\eta,x) \in R_1 - \partial\mathcal{E}_0$ *is erroneous, and so* $(\xi,\eta,x) \in R_1$ *implies* $(\xi,\eta,x) \in \partial\mathcal{E}_0$.

Similarly, $(\xi,\eta,x) \in R_2$ *implies* $(\xi,\eta,x) \in \partial\mathcal{E}_0$. ■

Since for any contact point (ξ,η), x such that $(\xi,\eta,x) \in \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2) \cap \partial\mathcal{E}_2$ has magnitude $\sqrt{1-\xi^2-\eta^2}$, and all x with this magnitude are such that $(\xi,\eta,x) \in \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2) \cap \partial\mathcal{E}_2$, (ξ,η) can be identified with the set $\{(\xi,\eta,x) \in \mathbb{R}^p \colon x^{\mathrm{T}}x = 1-\xi^2-\eta^2\}$. This means that when a contact point (ξ,η) is being mentioned, what in fact is being discussed is a typical member of the equivalence class $\{(\xi,\eta,x) \colon x^{\mathrm{T}}x = 1-\xi^2-\eta^2\}$.

Lemma 5.17: If V is open and

$$v^{\mathrm{T}}\epsilon \geq 0 \ \forall \epsilon \in V \ni 0 \text{ such that } {n_1}^{\mathrm{T}}\epsilon, \ldots, {n_\ell}^{\mathrm{T}}\epsilon < 0 \tag{5.87}$$

then, either one of the n_i is a non-positive linear combination of the others, in which case v is an arbitrary vector, or

$$v = k_1 n_1 + \cdots + k_\ell n_\ell, \ k_1, \ldots, k_\ell \leq 0. \tag{5.88}$$

□

Proof 5.17: *First, Gordan's theorem[20] will be quoted.*

If $A \in \mathbb{R}^{\ell \times p}$

$$Ax > 0 \tag{5.89}$$

has a solution $x \in \mathbb{R}^p$, *or*

$$A^{\mathrm{T}}y = 0, \ y \geq 0, \ y \neq 0 \tag{5.90}$$

has a solution $y \in \mathbb{R}^\ell$, *but never both.*

(*"*$>$*", "*$<$*", "*$\geq$*" and "*$\leq$*" for vectors will mean that the corresponding scalar relation holds for* each *of the components. For matrices, these symbols will have their usual meaning regarding positive and negative (semi)definiteness.)*

If $v^{\mathrm{T}}\epsilon \geq 0$ *for all* $\epsilon \in V \ni 0$ *such that* ${n_1}^{\mathrm{T}}\epsilon, \ldots, {n_\ell}^{\mathrm{T}}\epsilon < 0$, *then* $v^{\mathrm{T}}\epsilon \geq 0$ *for all* ϵ *such that* ${n_1}^{\mathrm{T}}\epsilon, \ldots, {n_\ell}^{\mathrm{T}}\epsilon < 0$. *This means that* $-[v, n_1, \ldots, n_\ell]^{\mathrm{T}}\epsilon > 0$ *has no solution* ϵ. *But then, by*

Gordan's theorem, there exists $k_0', k_1', \ldots, k_\ell' \leq 0$, *not all of which are zero, such that* $k_0' v + k_1' n_1 + \cdots + k_\ell' n_\ell = 0$. *If, for any such* $k_0', k_1', \ldots, k_\ell'$, $k_0' = 0$, *then one of the* n_i *is a non-positive linear combination of the others, which means that* $\{\epsilon : {n_i}^{\mathrm{T}} \epsilon < 0, i = 1, \ldots, \ell\} = \emptyset$ *and* v *is an arbitrary vector. If, for any such* $k_0', k_1', \ldots, k_\ell'$, $k_0' \neq 0$, *then there exists* $k_1, \ldots, k_\ell \leq 0$ *such that* $v = k_1 n_1 + \cdots + k_\ell n_\ell$. ■

Lemma 5.18: Suppose V is an open neighbourhood of 0, $A \in \mathbb{R}^{2\times 2}$ is symmetric and $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ for all $\epsilon \in V$, such that ${n_1}^{\mathrm{T}}\epsilon, \ldots, {n_\ell}^{\mathrm{T}}\epsilon < 0$, where the n_i are nonzero. Then, there are three possibilities:

there exists i_1 such that $n_i = k_i n_{i_1}$ for some $k_i > 0$, $i = 1, \ldots, \ell$ and then A must be positive semi-definite;

or there exist $i_1, i_2 \in \{1, \ldots, \ell\}$ such that $\{n_{i_1}, n_{i_2}\}$ is a linearly independent set and $n_i = k_{1i} n_{i_1} + k_{2i} n_{i_2}$, for some $k_{1i}, k_{2i} \geq 0$, and then $A = b_{11} n_{i_1} {n_{i_1}}^{\mathrm{T}} + b_{12}(n_{i_1} {n_{i_2}}^{\mathrm{T}} + n_{i_2} {n_{i_1}}^{\mathrm{T}}) + b_{22} n_{i_2} {n_{i_2}}^{\mathrm{T}}$ for b_{11}, b_{12}, b_{22} such that $b_{11}, b_{22} \geq 0$ and either $b_{11} b_{22} - b_{21}^2 \geq 0$ (when A is positive semi-definite) or $b_{12} \geq 0$;

or neither of the above hold, and then A is unrestricted.

□

Proof 5.18: *If* $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ *for all* $\epsilon \in V$ *such that* ${n_1}^{\mathrm{T}}\epsilon, \ldots, {n_\ell}^{\mathrm{T}}\epsilon < 0$, *then* $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ *for all* ϵ *such that* ${n_1}^{\mathrm{T}}\epsilon, \ldots, {n_\ell}^{\mathrm{T}}\epsilon < 0$.

Suppose there exists i_1 *such that there exists* $k_i > 0$ *such that* $n_i = k_i n_{i_1}$, $i = 1, \ldots, \ell$. *Then* ${n_1}^{\mathrm{T}}\epsilon, \ldots, {n_\ell}^{\mathrm{T}}\epsilon < 0$ *is equivalent to* ${n_{i_1}}^{\mathrm{T}}\epsilon < 0$. *Let* ϵ *be an arbitrary nonzero vector. If* ${n_{i_1}}^{\mathrm{T}}\epsilon < 0$, $\epsilon^{\mathrm{T}} A \epsilon$ *must not be less than zero. If* ${n_{i_1}}^{\mathrm{T}}\epsilon > 0$, *then* ${n_{i_1}}^{\mathrm{T}}(-\epsilon) < 0$, *so* $\epsilon^{\mathrm{T}} A \epsilon = (-\epsilon)^{\mathrm{T}} A (-\epsilon)$ *must not be less than zero. But, if* $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ *for* ϵ *such that* ${n_{i_1}}^{\mathrm{T}}\epsilon \neq 0$, $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ *for* ϵ *such that* ${n_{i_1}}^{\mathrm{T}}\epsilon = 0$, *by the continuity of* $\epsilon^{\mathrm{T}} A \epsilon$ *considered as a function of* ϵ. *Thus,* $\epsilon^{\mathrm{T}} A \epsilon$ *must not be less than zero for arbitrary* ϵ *and so* A *must be positive semi-definite.*

Suppose there exists i_1, i_2 *such that* $\{n_{i_1}, n_{i_2}\}$ *is linearly independent and there exists* $k_{1i}, k_{2i} \geq 0$ *such that* $n_i = k_{1i} n_{i_1} + k_{2i} n_{i_2}$, $i = 1, \ldots, \ell$. *Since the* n_i *are nonzero,* $(k_{1i}, k_{2i}) \neq 0$, $i = 1, \ldots, \ell$. *Suppose* ${n_{i_j}}^{\perp} \neq 0$ *is perpendicular to* n_{i_j}, $j = 1, 2$, *and that the choice of perpendicular directions is fixed by* ${n_{i_1}}^{\mathrm{T}} {n_{i_2}}^{\perp}, {n_{i_2}}^{\mathrm{T}} {n_{i_1}}^{\perp} > 0$ *(since* $\{n_{i_1}, n_{i_2}\} \subset \mathbb{R}^2$ *is linearly independent, so is* $\{{n_{i_1}}^{\perp}, {n_{i_2}}^{\perp}\}$, *and these inner products cannot be* 0*). Then, an arbitrary* ϵ *can be given by* $\epsilon = k_1 {n_{i_1}}^{\perp} + k_2 {n_{i_2}}^{\perp}$. *The requirements that* $\epsilon^{\mathrm{T}} n_{i_j} < 0$, $j = 1, 2$, *are satisfied by restricting* $k_i < 0$, $i = 1, 2$. *But then* $\epsilon^{\mathrm{T}} n_i = k_1 k_{2i} {n_{i_2}}^{\mathrm{T}} {n_{i_1}}^{\perp} + k_2 k_{1i} {n_{i_1}}^{\mathrm{T}} {n_{i_2}}^{\perp} < 0$, *so* ϵ *satisfies* all

the requirements on it. Thus A must satisfy $(k_1 {n_{i_1}}^{\perp} + k_2 {n_{i_2}}^{\perp})^{\mathrm{T}} A(k_1 {n_{i_1}}^{\perp} + k_2 {n_{i_2}}^{\perp}) \geq 0$ for all $k_1, k_2 < 0$. But $\{n_{i_1}, n_{i_2}\}$ is linearly independent, so there exist scalars b_{11}, b_{12}, b_{22} such that $A = b_{11} n_{i_1} {n_{i_1}}^{\mathrm{T}} + b_{12}(n_{i_1} {n_{i_2}}^{\mathrm{T}} + n_{i_2} {n_{i_1}}^{\mathrm{T}}) + b_{22} n_{i_2} {n_{i_2}}^{\mathrm{T}}$. Then $(k_1 {n_{i_1}}^{\perp} + k_2 {n_{i_2}}^{\perp})^{\mathrm{T}} A(k_1 {n_{i_1}}^{\perp} + k_2 {n_{i_2}}^{\perp}) = k_1^2 b_{22}({n_{i_2}}^{\mathrm{T}} {n_{i_1}}^{\perp})^2 + 2k_1 k_2 b_{12}({n_{i_2}}^{\mathrm{T}} {n_{i_1}}^{\perp})({n_{i_1}}^{\mathrm{T}} {n_{i_2}}^{\perp}) + k_2^2 b_{11}({n_{i_1}}^{\mathrm{T}} {n_{i_2}}^{\perp})^2$, and this should be nonnegative for all $k_1, k_2 < 0$. Bearing in mind that ${n_{i_2}}^{\mathrm{T}} {n_{i_1}}^{\perp}, {n_{i_1}}^{\mathrm{T}} {n_{i_2}}^{\perp} > 0$, this means that $b_{11}, b_{22} \geq 0$ and either $b_{11} b_{22} - b_{21}^2 \geq 0$, or $b_{12} \geq 0$.

Suppose the set $\{n_i, n_j\}$ is linearly dependent for some i, j such that $1 \leq i, j \leq \ell$ and there exist i_1, i_2 such that $n_{i_2} = k n_{i_1}$ where $k < 0$. Then ${n_{i_1}}^{\mathrm{T}} \epsilon < 0, {n_{i_2}}^{\mathrm{T}} \epsilon < 0$ is impossible, and so the condition $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ for all ϵ such that ${n_1}^{\mathrm{T}} \epsilon, \ldots, {n_\ell}^{\mathrm{T}} \epsilon < 0$ is empty and A is unrestricted.

Suppose no two of the n's are such that one is a negative multiple of the other and that there exists i, j such that the set $\{n_i, n_j\}$ is linearly independent, but suppose that for all such sets consisting of a pair of linearly independent vectors there exists n_m such that $n_m = k_i n_i + k_j n_j$ where at least one of k_i, k_j is negative.

With a suitable re-labelling if necessary, suppose $n_1, \ldots, n_m$ are such that $n_i = k_{m1i} n_1 + k_{mmi} n_m$, $k_{m1i}, k_{mmi} \geq 0$, $i = 1, \ldots, m$, and, if $i > m$, $\{n_1, n_i\}$ is linearly independent and $n_i = k_{m1i} n_1 + k_{mmi} n_m$, where at least one of k_{m1i}, k_{mmi} is negative. Since there exists at least one n_i such that $n_i = k_{m1i} n_1 + k_{mmi} n_m$, $k_{m1i}, k_{mmi} \geq 0$ does not hold, $m \neq \ell$.

Without loss of generality, suppose $k_{m1,m+1} < 0$ The case where $k_{mm,m+1} = 0$ (so $n_{m+1} = k_{m1,m+1} n_1$ for $k_{m1,m+1} < 0$) has already been dealt with, so it may be assumed that $k_{mm,m+1} \neq 0$. If $k_{mm,m+1} > 0$, then $n_m = -k_{mm,m+1}^{-1} k_{m1,m+1} n_1 + k_{mm,m+1}^{-1} n_{m+1}$ and $n_i = k_{m1i} n_1 + k_{mmi} n_m = \left(k_{m1i} - k_{mm,m+1}^{-1} k_{mmi} k_{m1,m+1}\right) n_1 + k_{mm,m+1} k_{mmi} n_{m+1} = k_{m+1,1i} n_1 + k_{m+1,m+1,i} n_{m+1}$, where $k_{m+1,1i}, k_{m+1,m+1,i} \geq 0$, $i = 1, \ldots, m+1$. (If any n_i, $i > m+1$, can be written as a nonnegative linear combination of n_1 and n_{m+1}, this process of writing as many of the n_i as possible as a nonnegative linear combination of a choice of two of the n_i can be speeded up by a further re-labelling).

If this process can be continued, the relations $n_i = k_{\ell 1i} n_1 + k_{\ell\ell i} n_\ell$, $k_{\ell 1i}, k_{\ell\ell i} \geq 0$, $i = 1, \ldots, \ell$ will be obtained, a contradiction. Thus there exists $m < \ell$ such that $n_{m+1} = k_{m1,m+1} n_1 + k_{mm,m+1} n_m$ and $k_{m1,m+1}, k_{mm,m+1} < 0$. But then ${n_1}^{\mathrm{T}} \epsilon, {n_m}^{\mathrm{T}} \epsilon, {n_{m+1}}^{\mathrm{T}} \epsilon < 0$ is impossible, so $\{\epsilon \colon {n_i}^{\mathrm{T}} \epsilon < 0, i = 1, \ldots, \ell\} = \emptyset$ and A is unrestricted. ■

Lemma 5.19: Suppose V is a neighbourhood of 0, $A \in \mathbb{R}^{2\times 2}$ and A is such that $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ for all $\epsilon \in V$ such that $n^{\mathrm{T}} \epsilon = 0$ and ${n_1}^{\mathrm{T}} \epsilon, \ldots, {n_\ell}^{\mathrm{T}} \epsilon < 0$, where n and the n_i are nonzero.

Let $n^{\perp}$ be one of the two unit vectors perpendicular to n. Then, there are two possibilities:

$n_i{}^{\mathrm{T}} n^{\perp} \neq 0, i = 1, \ldots, \ell$ and all these inner products have the same sign. In this case, only those A's such that $n^{\perp \mathrm{T}} A n^{\perp} \geq 0$ satisfy the above condition;

or this not the case. Then A is unrestricted.

□

Proof 5.19: *If A is such that $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ for all $\epsilon \in V$ such that $n^{\mathrm{T}} \epsilon = 0$ and $n_1{}^{\mathrm{T}} \epsilon, \ldots, n_\ell{}^{\mathrm{T}} \epsilon < 0$, then A is such that $\epsilon^{\mathrm{T}} A \epsilon \geq 0$ for all ϵ such that $n^{\mathrm{T}} \epsilon = 0$ and $n_1{}^{\mathrm{T}} \epsilon, \ldots, n_\ell{}^{\mathrm{T}} \epsilon < 0$. If $n^{\mathrm{T}} \epsilon = 0$, $\epsilon = \pm |\epsilon| n^{\perp}$. In the second case above, either there exists n_i such that $n_i{}^{\mathrm{T}} n^{\perp} = 0$, so $n_i{}^{\mathrm{T}} \epsilon = 0$, or there exists n_i, n_j such that $n_i{}^{\mathrm{T}} n^{\perp} > 0 > n_j{}^{\mathrm{T}} n^{\perp}$, so either $n_i{}^{\mathrm{T}} \epsilon > 0 > n_j{}^{\mathrm{T}} \epsilon$ or $n_j{}^{\mathrm{T}} \epsilon > 0 > n_i{}^{\mathrm{T}} \epsilon$, unless $\epsilon = 0$. Hence the set of appropriate ϵ is empty and A is unrestricted.*
In the first case in the statement of the Lemma, either $n_i{}^{\mathrm{T}} \epsilon < 0, i = 1, \ldots, \ell$ or $n_i{}^{\mathrm{T}} (-\epsilon) < 0, i = 1, \ldots, \ell$, so $|\epsilon|^2 n^{\perp \mathrm{T}} A n^{\perp} \geq 0$ implying $n^{\perp \mathrm{T}} A n^{\perp} \geq 0$. ■

Theorem 5.20: If (ξ, η) is a point of Type 0a, then the coincidence condition

$$\alpha(\xi - \xi_0)^2 + 2\gamma(\xi - \xi_0)(\eta - \eta_0) + \beta(\eta - \eta_0)^2 + \delta(1 - \xi^2 - \eta^2) = 1, \tag{5.91}$$

the tangency conditions

$$\begin{aligned} \alpha(\xi - \xi_0) + \gamma(\eta - \eta_0) &= \delta\xi \\ \gamma(\xi - \xi_0) + \beta(\eta - \eta_0) &= \delta\eta, \end{aligned} \tag{5.92}$$

and the Jacobian condition

$$\mathrm{I} - \delta^{-1} \bar{Q}^{-1} \geq 0 \tag{5.93}$$

hold.

If (ξ, η) is a point of Type 0b, then

$$\xi^2 + \eta^2 = 1, \tag{5.94}$$

the coincidence condition

$$\alpha(\xi - \xi_0)^2 + 2\gamma(\xi - \xi_0)(\eta - \eta_0) + \beta(\eta - \eta_0)^2 = 1, \tag{5.95}$$

and the tangency conditions

$$\begin{aligned} (k + \delta^{-1})[\alpha(\xi - \xi_0) + \gamma(\eta - \eta_0)] &= \xi \\ (k + \delta^{-1})[\gamma(\xi - \xi_0) + \beta(\eta - \eta_0)] &= \eta \end{aligned} \tag{5.96}$$

for some $k \leq 0$, hold.

If $\xi = \xi_0$ and $\eta = \eta_0$, or if $k = 0$ above, then inequality (5.93) holds in this case too, but, in any case, the weaker condition

$$(\delta - \alpha)\eta^2 + 2\gamma\eta\xi + (\delta - \beta)\xi^2 \geq 0 \tag{5.97}$$

holds. □

Proof 5.20: *If (ξ, η) is a contact point, then*

$$\alpha(\xi - \xi_0)^2 + 2\gamma(\xi - \xi_0)(\eta - \eta_0) + \beta(\eta - \eta_0)^2 + \delta x^{\mathrm{T}} x = 1 \tag{5.98}$$

for all points $x \in \mathbb{R}^{p-2}$ such that $\xi^2 + \eta^2 + x^{\mathrm{T}}x = 1$, so equation (5.91) must hold.
Let x_0 stand for the magnitude of an arbitrary x such that (ξ, η, x) is in the equivalence class of points represented by (ξ, η). Since (ξ, η) is a Type 0 point, there exists a neighbourhood V of (ξ, η, x) such that $V \cap (\mathbb{H}_1^+ \cup \mathbb{H}_1^- \cup \mathbb{H}_2^+ \cup \mathbb{H}_2^-) = \emptyset$. The equation

$$\begin{bmatrix} \xi - \xi_0 & \eta - \eta_0 \end{bmatrix} \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix} \begin{bmatrix} \xi - \xi_0 \\ \eta - \eta_0 \end{bmatrix} + \delta x_0^2 = 1, \tag{5.99}$$

which is $(\bar{x} - \bar{c})^{\mathrm{T}} \bar{Q}^{-1} (\bar{x} - \bar{c}) + \delta x_0^2 = 1$, can be regarded as implicitly defining nonnegative x_0 as a function of ξ and η. Then

$$\frac{\partial}{\partial \bar{x}^{\mathrm{T}}} x_0(\xi, \eta) = -\frac{1}{\delta x_0(\xi, \eta)} \bar{Q}^{-1}(\bar{x} - \bar{c}). \tag{5.100}$$

The distance of any point in the equivalence class represented by (ξ, η) from the centre of $\mathcal{E}_0$ is $\bar{x}^{\mathrm{T}}\bar{x} + x_0^2$, and

$$\frac{\partial}{\partial \bar{x}^{\mathrm{T}}}(\bar{x}^{\mathrm{T}}\bar{x} + x_0^2) = 2\bar{x} - 2\delta^{-1}\bar{Q}^{-1}(\bar{x} - \bar{c}), \tag{5.101}$$

$$\frac{\partial^2}{\partial \bar{x} \partial \bar{x}^{\mathrm{T}}}(\bar{x}^{\mathrm{T}}\bar{x} + x_0^2) = 2\mathrm{I} - 2\delta^{-1}\bar{Q}^{-1}. \tag{5.102}$$

Let $\epsilon_\xi, \epsilon_\eta$ be sufficiently small that

$$(\bar{x} + \epsilon - \bar{c})^{\mathrm{T}} \bar{Q}^{-1} (\bar{x} + \epsilon - \bar{c}) + \delta x_0(\bar{x} + \epsilon)^2 = 1 \tag{5.103}$$

has a real solution $x_0(\bar{x} + \epsilon)$ for $\epsilon = (\epsilon_\xi, \epsilon_\eta)$ (this will be possible as $\xi^2 + \eta^2 < 1$) and $\bar{x} + \epsilon \in \pi(V)$. Then $(\xi + \epsilon_\xi, \eta + \epsilon_\eta, x) \in \partial\mathcal{E}_2$ for all x such that $x^{\mathrm{T}}x = x_0(\bar{x} + \epsilon)^2$ and

$$\begin{aligned} (\bar{x} + \epsilon)^{\mathrm{T}}(\bar{x} + \epsilon) + x_0(\bar{x} + \epsilon)^2 &= \bar{x}^{\mathrm{T}}\bar{x} + x_0(\bar{x})^2 + \epsilon^{\mathrm{T}} \frac{\partial}{\partial \bar{x}^{\mathrm{T}}}(\bar{x}^{\mathrm{T}}\bar{x} + x_0(\bar{x})^2) + \\ &\qquad \frac{1}{2}\epsilon^{\mathrm{T}} \frac{\partial^2}{\partial \bar{x} \partial \bar{x}^{\mathrm{T}}}(\bar{x}^{\mathrm{T}}\bar{x} + x_0(\bar{x})^2)\epsilon \qquad (5.104) \\ &= 1 + 2\epsilon^{\mathrm{T}}(\bar{x} - \delta^{-1}\bar{Q}^{-1}(\bar{x} - \bar{c})) + \\ &\qquad \epsilon^{\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})\epsilon. \qquad (5.105) \end{aligned}$$

If $(\bar{x}+\epsilon)^{\mathrm{T}}(\bar{x}+\epsilon)+x_0(\bar{x}+\epsilon)^2 \geq 1$ *for all sufficiently small* ϵ*, which must be the case if no point of* $\partial\mathcal{E}_2 \cap V$ *is to be an interior point of* $\mathcal{E}_0$*, then* $\bar{x}-\delta^{-1}\bar{Q}^{-1}(\bar{x}-\bar{c})$ *must vanish and equation (5.92) must hold. In addition,* $(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})$ *must be positive semi-definite. That is, equation (5.93) must hold.*

For the case where (ξ,η) *is a point of Type 0b, derivation of the vector* $\bar{x}-\delta^{-1}\bar{Q}^{-1}(\bar{x}-\bar{c})$ *and the tensor* $(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})$ *follows either in the same way* after changing the dependent variable *or by a symmetry argument. However, the vector* ϵ *now has some restrictions on its direction as well as on its magnitude, as the quantity* $x_0(\xi,\eta)^2=\delta^{-1}\left[1-(\bar{x}+\epsilon-\bar{c})^{\mathrm{T}}\bar{Q}^{-1}(\bar{x}+\epsilon-\bar{c})\right]$ *is nonnegative for sufficiently small (but nonzero)* ϵ *only if* $\epsilon^{\mathrm{T}}\bar{Q}^{-1}(\bar{x}-\bar{c})<0$*. Set* $w=\bar{Q}^{-1}(\bar{x}-\bar{c})$*. Thus it is required that* $2\epsilon^{\mathrm{T}}(\bar{x}-\delta^{-1}w)+\epsilon^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})\epsilon\geq 0$ *for all* ϵ *such that* $\epsilon^{\mathrm{T}}w<0$ *and* ϵ *is in a neighbourhood* V *of 0. This means that there exists* $\epsilon_0\in\mathbb{R}$ *such that* $2\epsilon_1\hat{\epsilon}^{\mathrm{T}}(\bar{x}-\delta^{-1}w)+\epsilon_1^2\hat{\epsilon}^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})\hat{\epsilon}\geq 0$ *for all* $\epsilon_1\in(0,\epsilon_0]$ *and all* $\hat{\epsilon}$ *such that* $\hat{\epsilon}^{\mathrm{T}}\hat{\epsilon}=1$ *and* $\hat{\epsilon}^{\mathrm{T}}w<0$*. Thus,* $2\hat{\epsilon}^{\mathrm{T}}(\bar{x}-\delta^{-1}w)+\epsilon_1\hat{\epsilon}^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})\hat{\epsilon}\geq 0$ *for all* $\epsilon_1\in(0,\epsilon_0]$ *and all* $\hat{\epsilon}$ *such that* $\hat{\epsilon}^{\mathrm{T}}\hat{\epsilon}=1$ *and* $\hat{\epsilon}^{\mathrm{T}}w<0$ *and so* $2\hat{\epsilon}^{\mathrm{T}}(\bar{x}-\delta^{-1}w)\geq 0$ *for all* $\hat{\epsilon}$ *such that* $\hat{\epsilon}^{\mathrm{T}}w<0$*. Then, by Lemma 5.17,* $\bar{x}-\delta^{-1}w=kw$ *for some* $k\leq 0$*, and equation (5.96) has been derived.*

But now $2k\hat{\epsilon}^{\mathrm{T}}w+\epsilon_1\hat{\epsilon}^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})\hat{\epsilon}\geq 0$ *for all* $\epsilon_1\in(0,\epsilon_0]$ *and all* $\hat{\epsilon}$ *such that* $\hat{\epsilon}^{\mathrm{T}}\hat{\epsilon}=1$ *and* $\hat{\epsilon}^{\mathrm{T}}w<0$*. If* $w=0$ *(equivalent to* $\bar{x}=\bar{c}$*, as* $\bar{Q}>0$*) or* $k=0$*, then* $\mathrm{I}-\delta^{-1}\bar{Q}^{-1}\geq 0$ *by Lemma 5.18. Suppose* $w\neq 0$ *and let* $w^{\perp}$ *be one of the unit vectors perpendicular to* w*. Suppose* $w^{\perp\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})w^{\perp}<0$*. Let* $v=(1-\epsilon_2^2)^{\frac{1}{2}}w^{\perp}+\epsilon_2 w/\sqrt{w^{\mathrm{T}}w}$ *for some* $\epsilon_2\in(-1,0)$*. Then* $v^{\mathrm{T}}v=1$*,* $v^{\mathrm{T}}w<0$ *and* $v^{\mathrm{T}}w+\epsilon_1 v^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})v=\epsilon_2(w^{\mathrm{T}}w)^{\frac{1}{2}}+\epsilon_1[(1-\epsilon_2^2)(w^{\mathrm{T}}w)w^{\perp\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})w^{\perp}+2\sqrt{w^{\mathrm{T}}w}\epsilon_2(1-\epsilon_2^2)^{\frac{1}{2}}w^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})w^{\perp}+\epsilon_2^2 w^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})w]/(w^{\mathrm{T}}w)$ *which can be made negative by choosing* $|\epsilon_2|$ *sufficiently small. This contradiction means that* $w^{\perp\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})w^{\perp}\geq 0$*.*

But, if $w\neq 0$*,* w *is parallel to* $\bar{x}$ *and so any vector* $\bar{x}^{\perp}$ *orthogonal to* $\bar{x}$ *in* $\mathbb{R}^2$ *is parallel to* $w^{\perp}$*. This means that inequality (5.97) holds, as* $\bar{x}^{\perp\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})\bar{x}^{\perp}\geq 0$ *holds if* $w^{\perp\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})w^{\perp}\geq 0$ *does.* ■

Theorem 5.21: Suppose (ξ,η) is a point of Type 1. For $x\in\mathbb{R}^{p-2}$ such that $x^{\mathrm{T}}x=1-\xi^2-\eta^2$, $[\xi\ \eta\ x^{\mathrm{T}}]^{\mathrm{T}}$ belongs to $\mathbb{H}$, one of the hyperplanes $\mathbb{H}_1^+,\mathbb{H}_1^-,\mathbb{H}_2^+,\mathbb{H}_2^-$, by the definition of a Type 1 point. Suppose $\mathbb{H}$ is defined by $n^{\mathrm{T}}[\xi\ \eta\ x^{\mathrm{T}}]^{\mathrm{T}}=\zeta$, where $n=[n_1\ n_2\ 0^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^p$ is the outward normal to $\mathbb{H}$. Then

$$n_1\xi+n_2\eta=\zeta \tag{5.106}$$

and

if (ξ, η) is a point of Type 1a, then the coincidence condition

$$\alpha(\xi - \xi_0)^2 + 2\gamma(\xi - \xi_0)(\eta - \eta_0) + \beta(\eta - \eta_0)^2 + \delta(1 - \xi^2 - \eta^2) = 1, \tag{5.107}$$

and the tangency conditions

$$\begin{aligned} \xi - \delta^{-1}[\alpha(\xi - \xi_0) + \gamma(\eta - \eta_0)] &= kn_1, \\ \eta - \delta^{-1}[\gamma(\xi - \xi_0) + \beta(\eta - \eta_0)] &= kn_2, \end{aligned} \tag{5.108}$$

for some $k \leq 0$ hold.

If $k = 0$, the Jacobian condition

$$\mathrm{I} - \delta^{-1}\bar{Q}^{-1} \geq 0 \tag{5.109}$$

holds, but, if $k \neq 0$, the weaker Jacobian condition

$$(\delta - \alpha)n_2^2 + 2\gamma n_1 n_2 + (\delta - \beta)n_1^2 \tag{5.110}$$

holds;

if (ξ, η) is a point of Type 1b, then

$$\xi^2 + \eta^2 = 1, \tag{5.111}$$

by definition, and the coincidence condition

$$\alpha(\xi - \xi_0)^2 + 2\gamma(\xi - \xi_0)(\eta - \eta_0) + \beta(\eta - \eta_0)^2 = 1, \tag{5.112}$$

and tangency conditions

$$\begin{aligned} \xi - \delta^{-1}[\alpha(\xi - \xi_0) + \gamma(\eta - \eta_0)] &= k_1 n_1 + k_2[\alpha(\xi - \xi_0) + \gamma(\eta - \eta_0)], \\ \eta - \delta^{-1}[\gamma(\xi - \xi_0) + \beta(\eta - \eta_0)] &= k_1 n_2 + k_2[\gamma(\xi - \xi_0) + \beta(\eta - \eta_0)], \end{aligned} \tag{5.113}$$

for some non-positive k_1 and k_2, hold.

A Jacobian condition also holds. This has the form

$$\mathrm{I} - \delta^{-1}\bar{Q}^{-1} \geq 0 \tag{5.114}$$

if $k_1 = k_2 = 0$ and $\bar{Q}^{-1}(\bar{x} - \bar{c}) = kn$ for some $k > 0$.

If $k_1 = k_2 = 0$ and $\bar{Q}^{-1}(\bar{x} - \bar{c})$ is not proportional to n,

$$\begin{aligned} \mathrm{I} - \delta^{-1}\bar{Q}^{-1} &= b_{11} n n^{\mathrm{T}} + b_{12}(n(\bar{Q}^{-1}(\bar{x} - \bar{c}))^{\mathrm{T}} + (\bar{Q}^{-1}(\bar{x} - \bar{c}))n^{\mathrm{T}}) + \\ & \qquad b_{22}(\bar{Q}^{-1}(\bar{x} - \bar{c}))(\bar{Q}^{-1}(\bar{x} - \bar{c}))^{\mathrm{T}}, \end{aligned} \tag{5.115}$$

where $b_{11}, b_{22} \geq 0$ and either $b_{11}b_{22} - b_{12}^2 \geq 0$ or $b_{12} \geq 0$.

If $k_1 < 0$ and either $w = kn$ for some $k \geq 0$, or $k_2 = 0$, the Jacobian condition is

$$(\delta - \alpha)n_2^2 + 2\gamma n_1 n_2 + (\delta - \beta)n_1^2 \geq 0. \tag{5.116}$$

If $k_1 = 0$, the condition

$$(\delta - \alpha)\eta^2 + 2\gamma\xi\eta + (\delta - \beta)\xi^2 \geq 0 \tag{5.117}$$

holds.

Otherwise, $I - \delta^{-1}\bar{Q}^{-1}$ is unrestricted.

□

Proof 5.21: *Let $w = \bar{Q}^{-1}(\bar{x} - \bar{c})$. As in the proof of Theorem 5.20, (ξ, η) is a point of Type 1 if and only if the appropriate coincidence condition holds, $\epsilon^{\mathrm{T}}(\bar{x} - \delta^{-1}w) \geq 0$ for ϵ in a certain subset of a neighbourhood $V \ni [\xi, \eta, x^{\mathrm{T}}]^{\mathrm{T}}$, where the only hyperplane $\mathbb{H}_{1,2}^{\pm}$ that V intersects is $\mathbb{H}$, and $\epsilon^{\mathrm{T}}(I - \delta^{-1}\bar{Q}^{-1})\epsilon \geq 0$ for ϵ in a, in general, different subset of V. If (ξ, η) is of Type 1a, then only the ϵ such that $[\xi, \eta, x^{\mathrm{T}}]^{\mathrm{T}} + \epsilon \in \Pi$, where Π is the strip bounded by $\mathbb{H}$, are relevant. That is, the condition on $(\bar{x} - \delta^{-1}\bar{Q}^{-1})$ is that $\epsilon^{\mathrm{T}}(\bar{x} - \delta^{-1}w) \geq 0$ for all $\epsilon \in V$ such that $n^{\mathrm{T}}\epsilon < 0$. By Lemma 5.17, equations (5.108) hold.*

The condition on $\mathrm{I} - \delta^{-1}\bar{Q}^{-1}$ is that $2k\epsilon^{\mathrm{T}} + \epsilon^{\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})\epsilon \geq 0$ for all $\epsilon \in V$ such that $n^{\mathrm{T}}\epsilon < 0$. By an argument similar to that of Theorem 5.20, if $k = 0$,inequality (5.109) holds. On the other hand, if $k < 0$, $n^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n^{\perp} \geq 0$ for $n^{\perp}$ perpendicular to n, so inequality (5.110) follows.

If (ξ, η) is of Type 1b, then only the ϵ such that $n^{\mathrm{T}}\epsilon < 0$ and $\epsilon^{\mathrm{T}}w < 0$ are relevant for the consideration of the condition placed on $\bar{x} - \delta^{-1}w$, so, by Lemma 5.17, equations (5.113) hold (as $w = k'n$, $k' < 0$ would imply $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2 = \emptyset$).

The condition on $\mathrm{I} - \delta^{-1}\bar{Q}^{-1}$ is that $2\epsilon^{\mathrm{T}}(k_1 n + k_2 w) + \epsilon^{\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})\epsilon \geq 0$ for all ϵ such that $\epsilon^{\mathrm{T}}n, \epsilon^{\mathrm{T}}w < 0$.

As before, this implies $\epsilon^{\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})\epsilon \geq 0$ for all ϵ such that $\epsilon^{\mathrm{T}}n, \epsilon^{\mathrm{T}}w \leq 0$, $\epsilon^{\mathrm{T}}(k_1 n + k_2 w) = 0$.

There are four cases:

a $k_1 = k_2 = 0$. *In this case inequalities (5.114) and (5.115) follow from Lemma 5.18;*

b $w = kn$, *for some* $k > 0$, *or* $k_2 = 0$. *Then* $n^{\perp}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n^{\perp} \geq 0$, *where* $n^{\perp}$ *is perpendicular to* n, *and inequality (5.116) follows;*

c $k_1 = 0$. *Then* $w^{\perp}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})w^{\perp} \geq 0$, *where* $w^{\perp}$ *is perpendicular to* w, *and inequality (5.117) follows;*

d *otherwise* $\mathrm{I} - \delta^{-1}\bar{Q}^{-1}$ *is unrestricted.*

■

Theorem 5.22: Suppose (ξ, η) is a point of Type 2. For $x \in \mathbb{R}^{p-2}$ such that $x^{\mathrm{T}}x = 1 - \xi^2 - \eta^2$, $[\xi\ \eta\ x^{\mathrm{T}}]^{\mathrm{T}}$ belongs to $\mathbb{H}_1 \cap \mathbb{H}_2$, where $\mathbb{H}_1$ is one of the hyperplanes $\mathbb{H}_1^+, \mathbb{H}_1^-$, and $\mathbb{H}_2$ is one of the hyperplanes $\mathbb{H}_2^+, \mathbb{H}_2^-$, by the definition of a Type 2 point. Suppose $\mathbb{H}_i$ is defined by ${n_i}^{\mathrm{T}}[\xi\ \eta\ x^{\mathrm{T}}]^{\mathrm{T}} = \zeta_i$, where the $n_i = [n_{i1}\ n_{i2}\ 0^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^p$ are the outward normals to $\mathbb{H}_i$, $i = 1, 2$. Then

$$\begin{aligned} n_{11}\xi + n_{12}\eta &= \zeta_1 \\ n_{21}\xi + n_{22}\eta &= \zeta_2 \end{aligned} \tag{5.118}$$

and

if (ξ, η) is a point of Type 2a, then the coincidence condition

$$\alpha(\xi - \xi_0)^2 + 2\gamma(\xi - \xi_0)(\eta - \eta_0) + \beta(\eta - \eta_0)^2 + \delta(1 - \xi^2 - \eta^2) = 1, \tag{5.119}$$

the tangency conditions

$$\begin{aligned} \xi - \delta^{-1}[\alpha(\xi - \xi_0) + \gamma(\eta - \eta_0)] &= k_1 n_{11} + k_2 n_{21}, \\ \eta - \delta^{-1}[\gamma(\xi - \xi_0) + \beta(\eta - \eta_0)] &= k_1 n_{12} + k_2 n_{22}, \end{aligned} \tag{5.120}$$

for some $k_1, k_2 \leq 0$ hold.

If $k_1 = k_2 = 0$ (equivalent to $\bar{x} - \delta^{-1}\bar{Q}^{-1}(\bar{x} - \bar{c}) = 0$) the Jacobian condition

$$\mathrm{I} - \delta^{-1}\bar{Q}^{-1} = b_{11}n_1{n_1}^{\mathrm{T}} + b_{12}(n_1{n_2}^{\mathrm{T}} + n_2{n_1}^{\mathrm{T}}) + b_{22}n_2{n_2}^{\mathrm{T}}, \tag{5.121}$$

where $b_{11}, b_{22} \geq 0$ and either $b_{11}b_{22} - b_{12}^2 \geq 0$ (equivalent to $\mathrm{I} - \delta^{-1}\bar{Q}^{-1} \geq 0$), or $b_{12} \geq 0$, holds.

If $k_i = 0, i \in \{1, 2\}$, then

$$n_j^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n_j^{\perp} \geq 0, \tag{5.122}$$

where $n_j^{\perp}$ is either of the unit vectors perpendicular to n_j, and $j \in \{1,2\} - \{i\}$.

Otherwise, $\mathrm{I} - \delta^{-1}\bar{Q}^{-1}$ is unrestricted;

if (ξ, η) is a point of Type 2b, then

$$\xi^2 + \eta^2 = 1, \tag{5.123}$$

by the definition of a Type 2b point, and the conditions applying to points of Type 2a also hold here.

□

Proof 5.22: *Most of the proof here is similar to the proof of Theorem 5.21. It is only necessary to show that the expected tangency conditions for points of Type 2b are equivalent to those for Type 2a (then the Jacobian conditions for Type 2b will also be identical to those for Type 2a). The expected conditions are:*

$$\begin{aligned}
\xi - \delta^{-1}[\alpha(\xi - \xi_0) + \gamma(\eta - \eta_0)] &= k_1 n_{11} + k_2 n_{21} + \\
&\qquad k_3[\alpha(\xi - \xi_0) + \gamma(\eta - \eta_0)], \\
\eta - \delta^{-1}[\gamma(\xi - \xi_0) + \beta(\eta - \eta_0)] &= k_1 n_{12} + k_2 n_{22} + \\
&\qquad k_3[\gamma(\xi - \xi_0) + \beta(\eta - \eta_0)],
\end{aligned} \tag{5.124}$$

where k_1, k_2 and k_3 are non-positive. However, if $\bar{Q}^{-1}(\bar{x} - \bar{c})$ is not a positive linear combination of n_1 and n_2, then $\mathcal{E}_2 \not\supset \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, and, if $\bar{Q}^{-1}(\bar{x} - \bar{c})$ is a nonnegative linear combination of n_1 and n_2, the tangency conditions holding for a point of Type 2a follow from the above conditions. ■

Remark 1: If any of the k's in the tangency conditions of Theorems 5.21 and 5.22 above are zero, the equations governing a Type ia point become those governing a Type $(i-1)$a point, and the equations governing a Type ib point either become those governing a Type $(i-1)$b point or those governing a Type ia point. Similarly, if two k's are zero, a Type 1b point is governed by equations like those pertaining to a Type 0a point, and so on.

Remark 2: The possible existence of Type 2b points will be disregarded in what follows, as the ultimate purpose of the present text is to find minimum-volume ellipsoids in the context of parameter estimation with random noise, and, in this context, given infinite precision, Type 2b points will occur with zero probability. Given the high precision of modern computing, they will occur with vanishingly small probability.

Now some possible combinations of contact points will be eliminated.

Theorem 5.23: If (ξ, η) is a Type 0 point and $\mathcal{E}_2 \neq \mathcal{E}_0$, (ξ, η) is a Type 0b point. □

Proof 5.23: *Suppose there exists a Type 0a point. Let a proper orthogonal transformation O_1 be applied which takes the Type 0a point into $(\xi, 0)$, $\xi \in [0, 1)$. Then the coincidence condition becomes*

$$\alpha(\xi - \xi_0)^2 - 2\gamma\eta_0(\xi - \xi_0) + \beta\eta_0^2 + \delta(1 - \xi^2) = 0, \tag{5.125}$$

and the tangency conditions become

$$\begin{aligned} \delta\xi - \alpha(\xi - \xi_0) + \gamma\eta_0 &= 0, & (5.126)\\ -\gamma(\xi - \xi_0) + \beta\eta_0 &= 0. & (5.127) \end{aligned}$$

Now $-1 < \xi_0 < \xi < 1$, by Lemma 5.15. Thus, equations (5.125) to (5.127) can be solved for α, γ and δ:

$$\begin{aligned} \alpha &= \frac{\xi}{(\xi - \xi_0)(1 - \xi\xi_0)} + \frac{\beta\eta_0^2}{(\xi - \xi_0)^2}, & (5.128)\\ \gamma &= \frac{\beta\eta_0}{\xi - \xi_0}, & (5.129)\\ \delta &= \frac{1}{1 - \xi\xi_0}. & (5.130) \end{aligned}$$

Then $\det \bar{Q}^{-1} > 0$ implies that

$$\frac{\beta\xi}{(\xi - \xi_0)(1 - \xi\xi_0)} > 0, \tag{5.131}$$

which in turn implies that $\xi > 0$. Since $(\xi, 0)$ is a Type 0a point, $\mathrm{I}_2 - \delta^{-1}\bar{Q}^{-1} \geq 0$. Thus, $\delta \geq \beta$ and $\delta \geq \alpha$. The second of these relations means that

$$\frac{-\xi_0}{(\xi - \xi_0)(1 - \xi\xi_0)} \geq \frac{\beta\eta_0^2}{(\xi - \xi_0)^2}, \tag{5.132}$$

with the consequence that $\xi_0 < 0$, which means that $\delta \leq 1$.

Thus

$$\begin{aligned} \det Q^{-1} &= \delta^{p-2}(\alpha\beta - \gamma^2) \leq \frac{\delta^{p-2}\beta\xi}{(\xi - \xi_0)(1 - \xi\xi_0)} \leq \frac{\delta^{p-1}\xi}{(\xi - \xi_0)(1 - \xi\xi_0)}\\ &\leq \frac{\xi}{(\xi - \xi_0)(1 - \xi\xi_0)} \leq \frac{\xi}{(\xi - \xi_0)} \leq 1 \end{aligned} \tag{5.133}$$

as $\xi > 0 > \xi_0 > -1$. But this is a contradiction if $\mathcal{E}_2 \neq \mathcal{E}_0$. ■

Thus, it has been shown that Type 0a points do not occur and that Type 2b points obey the same conditions as Type 2a points. Effectively, the only Types that need to be considered are 0b, 1a, 1b and 2a.

Now some possible combinations of points of various Types can be eliminated.

Theorem 5.24: If $\mathcal{E}_2 \neq \mathcal{E}_0$, then there exists 0 or 1 points of Type 0b. □

Proof 5.24: *Suppose there are two distinct Type 0b points and that an orthogonal transformation O_1 has been applied so that they are taken into $(\xi_1, \pm\eta_1)$, where $\xi_1^2 + \eta_1^2 = 1$ and $\eta_1 > 0$. Then the coincidence equations become*

$$\alpha(\xi_1 - \xi_0)^2 + 2\gamma(\xi_1 - \xi_0)(\pm\eta_1 - \eta_0) + \beta(\pm\eta_1 - \eta_0)^2 = 1, \tag{5.134}$$

the tangency equations are

$$\begin{aligned} \delta\xi_1 - \alpha(\xi_1 - \xi_0) - \gamma(\pm\eta_1 - \eta_0) &= k_\pm[\alpha(\xi_1 - \xi_0) + \gamma(\pm\eta_1 - \eta_0)], & (5.135)\\ \pm\delta\eta_1 - \gamma(\xi_1 - \xi_0) - \beta(\pm\eta_1 - \eta_0) &= k_\pm[\gamma(\xi_1 - \xi_0) + \beta(\pm\eta_1 - \eta_0)]. & (5.136) \end{aligned}$$

Equations (5.135) and (5.136) combine to yield

$$\xi_1[\gamma(\xi_1 - \xi_0) + \beta(\pm\eta_1 - \eta_0)] = \pm\eta_1[\alpha(\xi_1 - \xi_0) + \gamma(\pm\eta_1 - \eta_0)]. \tag{5.137}$$

Then equations (5.134) and (5.137) result in

$$\begin{aligned} \alpha(\xi_1 - \xi_0)^2 - 2\gamma\eta_0(\xi_1 - \xi_0) + \beta(\eta_0^2 + 1 - \xi_1^2) &= 1, & (5.138)\\ \gamma(\xi_1 - \xi_0) - \beta\eta_0 &= 0, & (5.139)\\ \xi_1[\gamma(\xi_1 - \xi_0) - \beta\eta_0] &= \gamma\eta_1^2, & (5.140)\\ \beta\xi_1 &= \alpha(\xi_1 - \xi_0) + \gamma\eta_0 & (5.141) \end{aligned}$$

(as $\eta_1 \neq 0$).

Suppose $\xi_0 = \xi_1$. Then, by equation (5.139), $\eta_0 = 0$; by (5.140), $\gamma = 0$; by (5.141), $\xi_1 = 0$, so $\xi_0 = 0$; and, by (5.138), $\beta = 1$. But, by equation (5.136), $\delta - \beta = k_+\beta \leq 0$, so $\delta \leq 1$. Since equation (5.97) holds, $\delta - \alpha \geq 0$ and so $\beta = 1 \geq \delta \geq \alpha$, which means that $\det Q^{-1} = \delta^{p-2}\alpha\beta \leq 1$ which is a contradiction.

Suppose now that $\xi_0 \neq \xi_1$. Then the above equations allow the conclusion that $\gamma = \eta_0 = 0$ and

$$\alpha = \frac{\xi_1}{(\xi_1 - \xi_0)(1 - \xi_1\xi_0)}, \ \beta = \frac{1}{1 - \xi_1\xi_0}. \tag{5.142}$$

Once again, equations (5.136) and (5.97) mean that $\beta \geq \delta \geq \alpha > 0$. This chain of relations means that either $\xi_1 > 0 \geq \xi_0$ or $\xi_1 < 0 \leq \xi_0$. Also

$$\det Q^{-1} = \frac{\delta^{p-2}\xi_1}{(\xi_1 - \xi_0)(1 - \xi_1\xi_0)^2} \leq \frac{\xi_1}{(\xi_1 - \xi_0)(1 - \xi_1\xi_0)^p}. \tag{5.143}$$

But this is the same situation as encountered in Theorem 5.23, and so there cannot be two Type 0b points when $\xi_0 \neq \xi_1$ either. ■

Theorems 5.23 and 5.24 mean that the only possible connected subsets of $\pi(\partial\mathcal{E}_0 \cap \partial\mathcal{E}_2 \cap [\mathbb{H}_1^+ \cup \mathbb{H}_1^- \cup \mathbb{H}_2^+ \cup \mathbb{H}_2^-])$ of one Type are those of one point of Types 0b, 1a, 1b or 2, or sets containing infinitely many points of Type 1.

Theorem 5.25: Suppose $\mathbb{H}$ is one of the hyperplanes $\mathbb{H}_{1,2}^{\pm}$. Let $\mathbb{H} = \{x \in \mathbb{R}^p : [n^{\mathrm{T}}, 0]x = \zeta\}$, where $[n_1, n_2, 0]^{\mathrm{T}} = [n^{\mathrm{T}}, 0]^{\mathrm{T}} \in \mathbb{R}^p$ is the unit normal in the outward direction from the strip Π bounded by $\mathbb{H}$. If $\pi(\mathbb{H})$ contains a point of Type 1a and a further, distinct, contact point, $\pi(\mathbb{H})$ contains infinitely many points of Type 1a. If $\pi(\mathbb{H})$ contains a point Type 1b, then $\pi(\mathbb{H})$ contains either 0 or infinitely many points of Type 1a. If $\pi(\mathbb{H})$ contains infinitely many points of Type 1a, then

$$\bar{Q}^{-1} = \frac{1-\delta(1-\zeta^2-(n^{\perp\mathrm{T}}\bar{c})^2)}{(\zeta - n^{\mathrm{T}}\bar{c})^2} nn^{\mathrm{T}} + \frac{\delta(n^{\perp\mathrm{T}}\bar{c})}{\zeta - n^{\mathrm{T}}\bar{c}}(n^{\perp}n^{\mathrm{T}} + nn^{\perp\mathrm{T}}) + \delta n^{\perp}n^{\perp\mathrm{T}}, \tag{5.144}$$

$$\bar{c}^{\mathrm{T}}\bar{c} < 1 \tag{5.145}$$

$$n^{\mathrm{T}}\bar{c} < \zeta \tag{5.146}$$

$$\delta \in \left(0, \frac{1}{1-\zeta(n^{\mathrm{T}}\bar{c})}\right) \cap \left(0, \frac{1}{1-\zeta(n^{\mathrm{T}}\bar{c}) - \max_{x\in\pi(\partial\mathcal{E}\cap\mathbb{H}\cap\Pi_1\cap\Pi_2)}\{(n^{\perp\mathrm{T}}\bar{c})(n^{\perp\mathrm{T}}x)\}}\right) \tag{5.147}$$

$$n^{\perp\mathrm{T}}\bar{c} \neq 0 \tag{5.148}$$

or

$$\bar{Q}^{-1} = \frac{1-\delta(1-\zeta^2)}{(\zeta - n^{\mathrm{T}}\bar{c})^2} nn^{\mathrm{T}} + \delta n^{\perp}n^{\perp\mathrm{T}}, \tag{5.149}$$

$$\bar{c}^{\mathrm{T}}\bar{c} < 1 \tag{5.150}$$

$$n^{\mathrm{T}}\bar{c} < \zeta \tag{5.151}$$

$$\delta \in \left(0, \frac{1}{1-\zeta\xi_0}\right) \cap \left(0, \frac{1}{1-\zeta(n^{\mathrm{T}}\bar{c})}\right] \tag{5.152}$$

$$n^{\perp\mathrm{T}}\bar{c} = 0 \tag{5.153}$$

where $n^{\perp}$ is one of the two unit vectors perpendicular to n (in $\mathbb{R}^2$!), for some δ and $\bar{c}$, and then $\pi(\partial\mathcal{E}_0 \cap \mathbb{H} \cap \Pi_1 \cap \Pi_2) = \{x \in \pi(\mathbb{H} \cap \Pi_1 \cap \Pi_2) : x \text{ is of Type 1 or Type 2}\}$, and there are at most two points of Type 2 in this set. □

Proof 5.25:

Obviously, if the cardinal $\overline{\overline{\pi(\partial\mathcal{E}_0 \cap \mathbb{H} \cap \Pi_1 \cap \Pi_2)}} > 1$, $\overline{\overline{\pi(\partial\mathcal{E}_0 \cap \mathbb{H} \cap \Pi_1 \cap \Pi_2)}} = \mathbf{c}$.

Suppose an orthogonal transformation O_1 has been applied such that $n = [1,0]^{\mathrm{T}}$. Suppose $(\zeta, \eta_1) \in \pi(\mathbb{H})$ is a Type 1a point. Then

$$\begin{aligned} \alpha(\zeta-\xi_0)^2 + 2\gamma(\zeta-\xi_0)(\eta_1-\eta_0) + \beta(\eta_1-\eta_0)^2 + & \\ \delta(1-\zeta^2-\eta_1^2) &= 1, \qquad (5.154) \\ \zeta - \delta^{-1}[\alpha(\zeta-\xi_0) + \gamma(\eta_1-\eta_0)] &= k, \qquad (5.155) \\ \eta_1 - \delta^{-1}[\gamma(\zeta-\xi_0) + \beta(\eta_1-\eta_0)] &= 0, \qquad (5.156) \end{aligned}$$

for some $k \leq 0$, by Theorem 5.21.

Suppose (ζ, η_2) is a further contact point. Then

$$\begin{aligned} \alpha(\zeta-\xi_0)^2 + 2\gamma(\zeta-\xi_0)(\eta_2-\eta_0) + \beta(\eta_2-\eta_0)^2 + & \\ \delta(1-\zeta^2-\eta_2^2) &= 1, \qquad (5.157) \end{aligned}$$

where $\eta_2 \neq \eta_1$.

It will also be noted, by Lemma 5.15 that $\zeta > \xi_0$

The difference of equations (5.154) and (5.157) leads to

$$2\gamma(\zeta-\xi_0) + \beta(\eta_1+\eta_2-2\eta_0) - \delta(\eta_2+\eta_1) = 0. \qquad (5.158)$$

The matrix formed from the coefficients of β and γ in this equation and equation (5.156) is non-singular as $\xi_0 \neq \zeta$ and $\eta_2 \neq \eta_1$, so these equations can be solved simultaneously for β and γ. The result of doing this can be substituted back into equation (5.154) and it can then be concluded that

$$\begin{aligned} \alpha &= \frac{1-\delta(1-\zeta^2-\eta_0^2)}{(\zeta-\xi_0)^2}, \qquad (5.159) \\ \beta &= \delta, \qquad (5.160) \\ \gamma &= \frac{\delta\eta_0}{\zeta-\xi_0}. \qquad (5.161) \end{aligned}$$

With these values of α,β and γ, $\alpha(\zeta-\xi_0)^2 + 2\gamma(\zeta-\xi_0)(\eta-\eta_0) + \beta(\eta-\eta_0)^2 + \delta(1-\zeta^2-\eta^2) = 1$ for all $\eta \in [-\sqrt{1-\zeta^2}, \sqrt{1-\zeta^2}]$, so all of these points are Type 1 points provided that $(\zeta, \eta) \in \pi(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$. But, obviously, if $\eta \in (\min\{\eta_1, \eta_2\}, \max\{\eta_1, \eta_2\})$, this condition is satisfied.

Then, either

$$\begin{aligned} \delta \in & \left(0, \frac{1}{1-\zeta^2}\right) \cap \\ & \left(0, \frac{1}{1-\zeta\xi_0 - \max\{\eta_0\eta \colon (\zeta,\eta) \in \pi(\partial\mathcal{E}_0 \cap \mathbb{H} \cap \Pi_1 \cap \Pi_2)\}}\right) \qquad (5.162) \end{aligned}$$

$$\eta_0 \neq 0 \qquad (5.163)$$

or

$$\alpha = \frac{1-\delta(1-\zeta^2)}{(\zeta-\xi_0)^2}, \tag{5.164}$$

$$\beta = \delta, \tag{5.165}$$

$$\gamma = 0 \tag{5.166}$$

$$\delta \in \left(0, \frac{1}{1-\zeta^2}\right) \cap \left(0, \frac{1}{1-\zeta\xi_0}\right]$$

$$\eta_0 = 0. \tag{5.167}$$

The restrictions on δ in equations (5.162) and (5.167) are necessary to ensure that an equation like (5.155) holds for at least two values of η and that $\alpha\beta - \gamma^2 > 0$ holds.

It is now desirable to write the relations (5.159) to (5.163, and relations (5.164) to (5.167), which ensure that $\partial\mathcal{E}_0 \cap \mathbb{H} = \{x \in \mathbb{H} : x$ is of Type 1$\}$, in a form which is unaffected by rotations O_1 which affect only e_1 and e_2. This is done by noting that $\alpha = n^{\mathrm{T}}\bar{Q}^{-1}n$, $\beta = n^{\perp\mathrm{T}}\bar{Q}^{-1}n^{\perp}$, $\gamma = n^{\mathrm{T}}\bar{Q}^{-1}n^{\perp}$, $\xi_0 = n^{\mathrm{T}}\bar{c}$ and $\eta_0 = \pm n^{\perp\mathrm{T}}\bar{c}$, where $n^{\perp}$ is one of the two unit vectors perpendicular to n. This, together with the facts that $\bar{Q}^{-1} = (n^{\mathrm{T}}\bar{Q}^{-1}n)nn^{\mathrm{T}} + (n^{\mathrm{T}}\bar{Q}^{-1}n^{\perp})(nn^{\perp\mathrm{T}} + n^{\perp}n^{\mathrm{T}}) + (n^{\perp\mathrm{T}}\bar{Q}^{-1}n^{\perp})n^{\perp}n^{\perp\mathrm{T}}$ and $\bar{c} = (n^{\mathrm{T}}\bar{c})n + (n^{\perp\mathrm{T}}\bar{c})n^{\perp}$ means that relations (5.159) to (5.163), or (5.164) to (5.167), which ensure that $\partial\mathcal{E}_0 \cap \mathbb{H} = \{x \in \mathbb{H}: x$ is of Type 1$\}$, can be written as relations (5.144) to (5.148), or (5.149) to (5.153) respectively. ■

Now some sets can be defined.

***Definition 5.5**: Let S be a connected component of $\pi(\partial\mathcal{E}_0 \cap \partial\mathcal{E}_2 \cap [\mathbb{H}_1^+ \cup \mathbb{H}_1^- \cup \mathbb{H}_2^+ \cup \mathbb{H}_2^-])$. If S consists of a single point of Types 0b, 1a, 1b or 2, S will be described as being T0b, T1a, T1b or T2 respectively. If S has infinitely many points, the sets $S \cap \pi(\mathbb{H})$, where $\mathbb{H} \in \{\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-\}$, will be described as being $\mathbb{H}T$ if they consist of more than one point.* □

A slight diversion will be made here to investigate what happens when $\mathbb{H}$ contains two points of Type 1b.

Theorem 5.26: Let the definitions of Theorem 5.25 hold. If $\mathbb{H}$ contains two points of Type 1b,

$$\bar{Q}^{-1} = \frac{1-\beta(1-\zeta^2-(n^{\perp\mathrm{T}}\bar{c})^2)}{(\zeta-n^{\mathrm{T}}\bar{c})^2}nn^{\mathrm{T}} + \frac{\beta n^{\perp\mathrm{T}}\bar{c}}{\zeta-n^{\mathrm{T}}\bar{c}}(n^{\perp}n^{\mathrm{T}} + nn^{\perp\mathrm{T}}) + \beta n^{\perp}n^{\perp\mathrm{T}}, \tag{5.168}$$

$$n^{\mathrm{T}}\bar{c} < \zeta, \tag{5.169}$$

$$\bar{c}^{\mathrm{T}}\bar{c} < 1, \tag{5.170}$$

$$\beta \in \left(0, \frac{1}{1-\zeta^2-(n^{\perp\mathrm{T}}\bar{c})^2}\right), \tag{5.171}$$

$$\delta \in (0, \beta], \tag{5.172}$$

for some $\bar{c} \in \mathbb{R}^2$, $\beta, \delta \in \mathbb{R}$. □

Proof 5.26: *Now suppose* (ζ, η_1) *and* $(\zeta, -\eta_1)$ *are distinct Type 1b points (and* $\eta_1 > 0$*). Then the coincidence conditions*

$$\alpha(\zeta - \xi_0)^2 + 2\gamma(\zeta - \xi_0)(\pm\eta_1 - \eta_0) + \beta(\pm\eta_1 - \eta_0)^2 = 1, \tag{5.173}$$

hold, which means that

$$\begin{aligned} \alpha(\zeta - \xi_0)^2 - 2\gamma\eta_0(\zeta - \xi_0) + \beta(1 - \xi_1^2 + \eta_0^2) &= 1, & (5.174)\\ 2\gamma\eta_1(\zeta - \xi_0) - 2\beta\eta_1\eta_0 &= 0. & (5.175) \end{aligned}$$

Since $\xi_0 < \zeta$ *and* $\eta_1 > 0$*, these equations can be solved for* α *and* γ*:*

$$\begin{aligned} \alpha &= \frac{1 - \beta(1 - \zeta^2 - \eta_0^2)}{(\zeta - \xi_0)^2}; & (5.176)\\ \gamma &= \frac{\beta\eta_0}{\zeta - \xi_0}. & (5.177) \end{aligned}$$

The second components of the tangency conditions are

$$\pm\delta\eta_1 - \gamma(\zeta - \xi_0) - \beta(\pm\eta_1 - \eta_0) = k_{2\pm}[\gamma(\zeta - \xi_0) + \beta(\pm\eta_1 - \eta_0)], \tag{5.178}$$

where $k_{2\pm} \leq 0$*, or*

$$\pm\delta\eta_1 \mp \beta\eta_1 = \pm k_{2\pm}\beta\eta_1 \Rightarrow \delta - \beta = k_{2\pm}\beta \leq 0 \tag{5.179}$$

so $k_{2+} = k_{2-} = \beta^{-1}\delta - 1$ *and* $\delta \leq \beta$*. This, together with equations (5.176) and (5.177) and Lemma 5.15 lead to equations (5.168) to (5.172) by the methods of the previous Theorem.* ■

The task of eliminating combinations of contact points will now be returned to.

Theorem 5.27: If $\mathbb{H} \in \{\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-\}$ is such that $\pi(\mathbb{H})$ contains two Type 1b points or infinitely many Type 1 points, then the opposite hyperplane $\mathbb{H}'$ contains 0 or infinitely many points of Type 1a. If $\mathbb{H}'$ contains infinitely many Type 1 points, then so does $\mathbb{H}$. □

Proof 5.27: *Suppose* $\mathbb{H} = \pi^{-1}(\{x \in \mathbb{R}^2 : n^{\mathrm{T}}x = \zeta\})$, $\mathbb{H}' = \pi^{-1}(\{x \in \mathbb{R}^2 : n^{\mathrm{T}}x = \zeta'\})$*, where* $[n^{\mathrm{T}}, 0]^{\mathrm{T}}$ *is the outward normal to the strip bounded by* $\mathbb{H}$ *and* $\mathbb{H}'$*. Suppose* n *has been taken into* $[1, 0]^{\mathrm{T}}$ *by an orthogonal transformation and that there is a Type 1a point on* $\pi(\mathbb{H}')$ *given by* (ζ', η)*. If* $\mathbb{H}$ *contains infinitely many points of Type 1, then* $\beta = \delta$*, by Theorem 5.25. If* $\mathbb{H}$ *contains two Type 1b points,* $\beta \geq \delta$*, by Theorem 5.26. But* $\beta \leq \delta$*, by Theorem 5.21 and the existence of a Type 1a point on* $\mathbb{H}'$*, so* $\beta = \delta$ *in this case too. Consequently, the second component of the tangency condition for the Type 1a point* (ζ', η) *is*

$$\delta\eta_0 - \gamma(\zeta' - \xi_0) = 0, \tag{5.180}$$

so

$$\gamma = \frac{\delta\eta_0}{\zeta' - \xi_0} = \frac{\delta\eta_0}{\zeta - \xi_0}, \tag{5.181}$$

by Theorem 5.25. But, of course, $\zeta' \neq \zeta$, *so* $\eta_0 = 0$ *and* $\gamma = 0$. *This implies that every point* $(\zeta', \eta') \in \pi(\mathbb{H}') \cup \pi(\mathbb{H}')$ *is a Type 1 point.* ■

Theorem 5.28: Let the conditions and definitions of Theorem 5.25 hold. If infinitely many points of Type 1a are contained in $\mathbb{H}$ and a further, distinct, contact point (ξ_3, η_3) is known, then (ξ_3, η_3) cannot be of Type 0b. □

Proof 5.28: *Let the orthogonal transformation of Theorem 5.25 be applied, if necessary. Then, clearly,* $\xi_3 < \zeta$.

Suppose (ξ_3, η_3) *is a Type 0b point. Then the associated coincidence condition is, by Theorem 5.20 and equations (5.159) to (5.163) or equations (5.164) to (5.167):*

$$\frac{1 - \delta(1 - \zeta^2 - \eta_0^2)}{(\zeta - \xi_0)^2}(\xi_3 - \xi_0)^2 + 2\frac{\delta\eta_0}{\zeta - \xi_0}(\xi_3 - \xi_0)(\eta_3 - \eta_0) + \delta(1 - \xi_3^2 - 2\eta_3\eta_0 + \eta_0^2) = 1, \tag{5.182}$$

and the tangency condition is

$$\begin{bmatrix} \delta\xi_3 - \frac{1-\delta(1-\zeta^2-\eta_0^2)}{(\zeta-\xi_0)^2}(\xi_3 - \xi_0) - \frac{\delta\eta_0}{\zeta-\xi_0}(\eta_3 - \eta_0) \\ \delta\eta_0 - \frac{\delta\eta_0}{\zeta-\xi_0}(\xi_3 - \xi_0) \end{bmatrix} =$$

$$k_3 \begin{bmatrix} \frac{1-\delta(1-\zeta^2-\eta_0^2)}{(\zeta-\xi_0)^2}(\xi_3 - \xi_0) + \frac{\delta\eta_0}{\zeta-\xi_0}(\eta_3 - \eta_0) \\ \frac{\delta\eta_0}{\zeta-\xi_0}(\xi_3 - \xi_0) + \delta(\eta_3 - \eta_0) \end{bmatrix}, \tag{5.183}$$

for $k_3 \leq 0$.

Eliminating k_3 *between the first and second components of equation (5.183) yields*

$$\begin{aligned} -\eta_0(\zeta - \xi_0)\delta - \eta_0\zeta(\zeta - \xi_0)\delta\xi_3 + [\xi_0 - ((1 - \zeta^2)\xi_0 - \zeta\eta_0^2)\delta]\eta_3 + \\ 2\eta_0(\zeta - \xi_0)\delta\xi_3^2 - [1 - (1 - 2\zeta\xi_0 + \xi_0^2 - \eta_0^2)\delta]\xi_3\eta_3 \quad = \quad 0. \end{aligned} \tag{5.184}$$

Suppose $\eta_0 = 0$. *Then equation (5.182) becomes*

$$[\xi_3 + \zeta - 2(1 + \xi_3\zeta)\xi_0 + (\xi_3 + \zeta)\xi_0^2]\delta = \xi_3 + \zeta - 2\xi 0. \tag{5.185}$$

If the coefficient of δ *in this equation is 0,* $\xi_0 = \frac{1}{2}(\xi_3 + \zeta)$, *and the coefficient of* δ *becomes* $\frac{1}{4}(\zeta^2 - \xi_3^2)$, *which can only vanish if* $\xi_3 = -\zeta$. *But, then* $\xi_0 = 0$, $\eta_3 = \pm\sqrt{1 - \zeta^2}$ *and equation (5.184) becomes* $\pm(\delta - 1)\zeta\sqrt{1 - \zeta^2} = 0$, *so* $\delta = 1$. *But, substituting this, together with* $\xi_0 = 0$ *in equations (5.164) to (5.167) would lead to the conclusion that* $\mathcal{E}_2 = \mathcal{E}_0$.

Now assume that the coefficient of δ in equation (5.185) does not vanish (a consequence of this is that $\xi_0 \neq \frac{1}{2}(\xi_3 + \zeta)$). Then that equation can be solved for δ:

$$\delta = \frac{2\xi_0 - \xi_3 - \zeta}{\xi_3 + \zeta - 2(1 + \xi_3\zeta)\xi_0 + (\xi_3 + \zeta)\xi_0^2}. \tag{5.186}$$

When this is substituted into equation (5.184), the result is $\eta_3(2\xi_0 - \xi_3 - \zeta)(\zeta - \xi_3)\xi_0 = 0$, and the only solution of this that is consistent with the assumptions in force is $\xi_0 = 0$. But, when this is put into the equation above, the result is $\delta = 1$, and, by equations (5.164) to (5.167), $\mathcal{E}_2 = \mathcal{E}_0$. Thus, if $\eta_0 = 0$, the contradiction $\mathcal{E}_2 = \mathcal{E}_0$ follows. Hence, $\eta_0 \neq 0$, and it may be assumed without loss of generality that $\eta_0 > 0$.

Then, equations (5.182) and (5.184) can be solved for ξ_3 and η_3, as (5.182) is linear in η_3 (with nonzero coefficient for η_3), and the result of substitituting the solution of (5.182) for η_3 in (5.184) has a numerator which is the product of two factors which are linear in ξ_3, one of which is the positive quantity $\zeta - \xi_3$. The other factor has a nonzero coefficient for ξ_3, as a consequence of the fact that $\xi_0 \neq \zeta$. When the result of solving (5.182) and (5.184) for ξ_3 and η_3 is substituted back into the equation $\xi_3^2 + \eta_3^2 = 1$, the result is $4\delta^2\eta_0^2(\zeta - \xi_0)^2$ times

$$\begin{aligned}
&(1 - \zeta^2)(1 - \xi_0^2 - \eta_0^2)\delta^2 + \\
&2[1 - \zeta^2 + 2\zeta\xi_0 - (3 + \zeta^2)\xi_0^2 - (1 + \zeta^2)\eta_0^2 + 2\zeta\xi_0(\xi_0^2 + \eta_0^2)]\delta + \\
&1 - (2\xi_0 - \zeta)^2 \;=\; 0.
\end{aligned} \tag{5.187}$$

As the other factors do not vanish, the solutions of this equation determine δ:

$$\begin{aligned}
\delta \;&=\; \frac{\begin{array}{r}[1 - \zeta^2 + 2\zeta\xi_0 - (3 + \zeta^2)\xi_0^2 + 2\zeta\xi_0^3 - (1 + \zeta^2 - 2\zeta\xi_0)\eta_0^2 \\ \pm 2(\zeta - \xi_0)\sqrt{(\zeta - \xi_0)^2\xi_0^2 - (1 - \zeta^2 + 2\zeta\xi_0 - 2\xi_0^2)\eta_0^2 + \eta_0^4}]\end{array}}{(1 - \zeta^2)(1 - \xi_0^2 - \eta_0^2)^2} \\
&=:\; \delta_\pm(\xi_0, \eta_0).
\end{aligned} \tag{5.188}$$

The functions $\delta_\pm$ are continuous on the set $\{(\xi_0, \eta_0) : \xi_0 \in (-1, \zeta), \eta_0 \in (0, \sqrt{1 - \xi_0^2})\}$ and they are real when the quantity under the square root in the above equation is nonnegative. This is the case when $\xi_0 \in ((-\infty, -\frac{1}{2}(1 - \zeta)] \cup [\frac{1}{2}(1 + \zeta), \infty)) \cap (-1, \zeta) = (-1, -\frac{1}{2}(1 - \zeta)]$ and when $\eta_0^2 \leq \eta_{r-}^{(2)}(\xi_0)$ or $\eta_0^2 \geq \eta_{r+}^{(2)}(\xi_0)$ for $\xi_0 \in (-\frac{1}{2}(1 - \zeta), \zeta)$ and

$$\eta_{r\pm}^{(2)}(\xi_0) := \xi_0(\zeta - \xi_0) + \frac{1 - \zeta^2 \pm \sqrt{(1 - \zeta^2)[1 - (2\xi_0 - \zeta)^2]}}{2}. \tag{5.189}$$

Let

$$\mathcal{S}_0 = \left\{(\xi,\eta)\colon \xi \in (-1,\zeta), \eta \in \left(0, \sqrt{1-\xi_0^2}\right)\right\}, \tag{5.190}$$

$$\mathcal{S}_1 = \left\{(\xi,\eta)\colon \xi \in \left(-\frac{1}{2}(1-\zeta), \frac{1}{2}(1+\zeta)\right), \eta \in \left(\sqrt{\eta_{r-}^{(2)}(\xi_0)}, \sqrt{\eta_{r+}^{(2)}(\xi_0)}\right)\right\}, \tag{5.191}$$

$$\mathcal{S}_2 = \mathcal{S}_0 - \mathcal{S}_1. \tag{5.192}$$

The equation $\delta_+(\xi_0,\eta_0) = 0$ *has the solutions* $\xi_0 = \pm\frac{1}{2}(1 \pm \zeta)$. *The value of* δ_+ *on the upper part of the boundary of* $\mathcal{S}_1$ *is*

$$\delta_+(\xi_0, \eta_{r+}^{(2)}(\xi_0)) = -\sqrt{1-(2\xi_0-\zeta)^2} \times \frac{1+\zeta^2-2\zeta\xi_0+\sqrt{(1-\zeta^2)[1-(2\xi_0-\zeta)^2]}}{(\zeta-\xi_0)^2\sqrt{1-\zeta^2}}. \tag{5.193}$$

This quantity is negative, and, moreover, $\delta_-(\xi_0, \eta_{r+}^{(2)}(\xi_0))$ *is clearly the same function of* ξ_0. *By the continuity of* $\delta_\pm$, $\delta_\pm(\xi_0,\eta_0) < 0$ *on the set*

$$\left\{(\xi_0,\eta_0)\colon \xi_0 \in \left(-\frac{1}{2}(1-\zeta), \zeta\right), \eta_0 \in \left(\sqrt{\eta_{r+}^{(2)}(\xi_0)}, \sqrt{1-\xi_0^2}\right)\right\}. \tag{5.194}$$

Along $\eta_0 = 0$, δ_+ *has the value*

$$\delta_+(\xi_0,0) = \begin{cases} \dfrac{1-\zeta+2\xi_0}{(1+\xi_0)^2(1-\zeta)}, & \text{if } \xi_0 \geq 0; \\ \dfrac{1+\zeta-2\xi_0}{(1-\xi_0)^2(1+\zeta)}, & \text{if } \xi_0 \leq 0, \end{cases} \tag{5.195}$$

Thus, $\delta_+(\xi_0,0) > 0$ *for* $\xi_0 \in \left(-1, -\frac{1}{2}(1-\zeta)\right) \cup \left(-\frac{1}{2}(1-\zeta), \zeta\right)$, *and so, by the continuity of* δ_+, $\delta_+(\xi_0,\eta_0) > 0$ *on the set*

$$\left\{(\xi_0,\eta_0)\colon \xi_0 \in \left(-1, -\frac{1}{2}(1-\zeta)\right), \eta_0 \in \left(0, \sqrt{1-\xi_0^2}\right)\right\} \cup \left\{(\xi_0,\eta_0)\colon \xi_0 \in \left(-\frac{1}{2}(1-\zeta), \zeta\right), \eta_0 \in \left(0, \sqrt{\eta_{r-}^{(2)}(\xi_0)}\right)\right\}. \tag{5.196}$$

Along $\eta_0 = 0$, δ_- *has the value*

$$\delta_-(\xi_0,0) = \begin{cases} \dfrac{1+\zeta-2\xi_0}{(1-\xi_0)^2(1+\zeta)}, & \text{if } \xi_0 \geq 0; \\ \dfrac{1-\zeta+2\xi_0}{(1+\xi_0)^2(1-\zeta)}, & \text{if } \xi_0 \leq 0, \end{cases} \tag{5.197}$$

Thus, $\delta_-(\xi_0, 0) > 0$ *for* $\xi_0 \in \left(-\frac{1}{2}(1-\zeta), \zeta\right)$*, and so, by the continuity of* δ_-*,* $\delta_-(\xi_0, \eta_0) > 0$ *on the set*

$$\left\{(\xi_0, \eta_0) : \xi_0 \in \left(-\frac{1}{2}(1-\zeta), \zeta\right), \eta_0 \in \left(0, \sqrt{\eta_{r-}^{(2)}(\xi_0)}\right)\right\}. \tag{5.198}$$

The regions where $\delta_\pm$ *are real and positive are illustrated in Figure 5.2.*

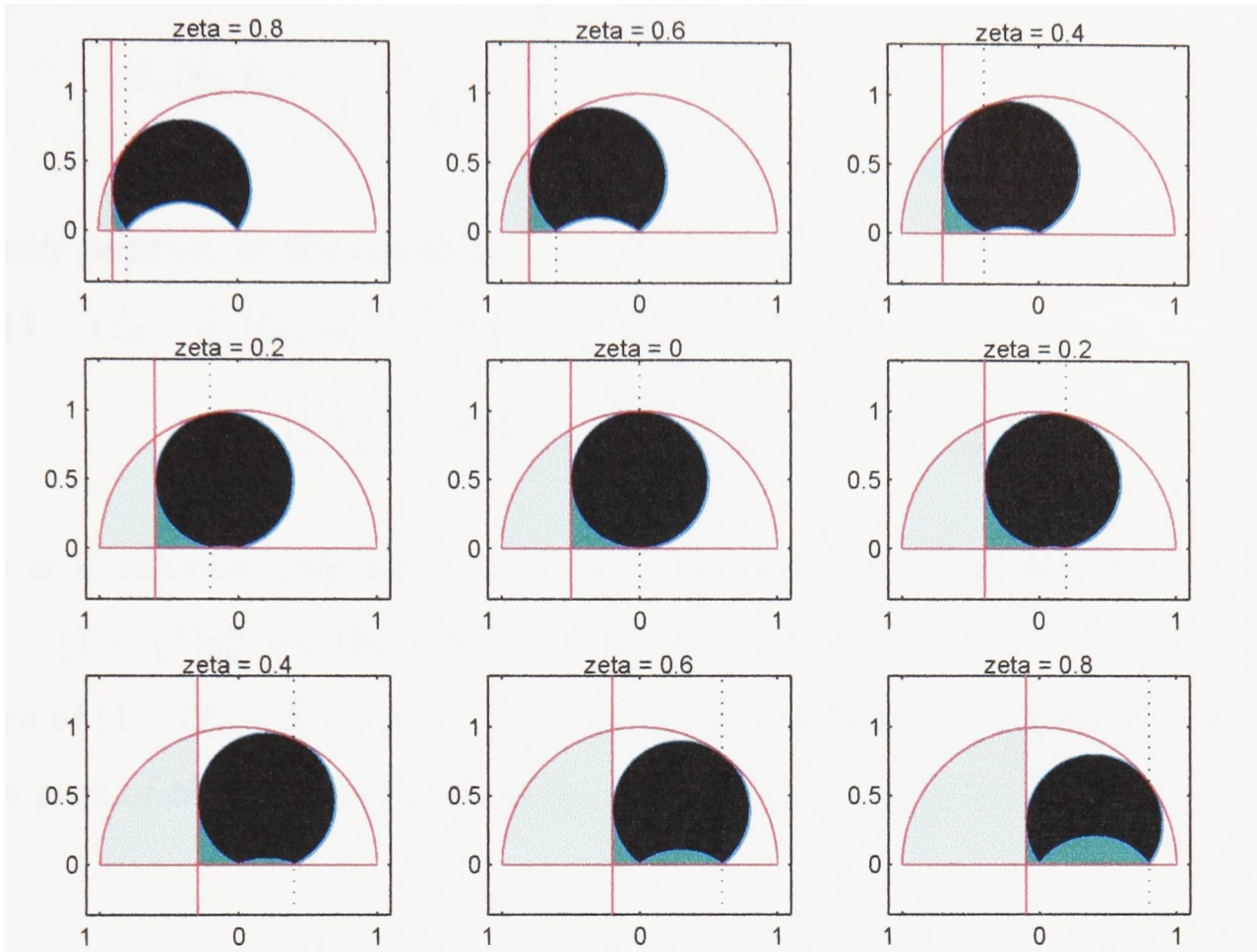


Figure 5.2: Regions of $\mathcal{S}_0$ where $\delta_\pm$ are real and positive. The set $\mathcal{S}_1$ is black, the region where both $\delta_\pm$ are real and positive is dark green, and the region where δ_+ alone is real and positive is light green. The curves which are to be included in those regions of positive, real $\delta_\pm$ whose boundary they intersect are cyan, and those which are to be excluded are red.

The requirement that $\delta < 1/(1 - \zeta\xi_0 - \max\{\eta_0\eta : (\zeta, \eta) \in \pi(\partial\mathcal{E}_0 \cap \mathbb{H} \cap \Pi_1 \cap \Pi_2)\})$ *implies that* $\delta < 1/(1 - \zeta\xi_0 - \sqrt{1-\zeta^2}\eta_0)$ *if* $\eta_0 > 0$*. This condition is equivalent to at least one of the following conditions holding:*

1. $\delta < 1/(1 - \zeta\xi_0)$;
2. $(1 - \zeta\xi_0 - \delta^{-1})^2 < \eta_0^2(1 - \zeta^2)$.

The equation $\delta_+(\xi_0, \eta_0) = 1/(1 - \zeta\xi_0)$ *only has the solutions*

$$\eta_0^2 = -\frac{(\zeta - \xi_0)[\sqrt{\zeta(1-\zeta\xi_0)} \pm \sqrt{\zeta - \xi_0}]^2}{1 - \zeta^2} \tag{5.199}$$

for η_0^2, but, as these are both negative, they can be disregarded as they do not lead to real solutions for η_0. The function $\delta_+(\xi_0,\eta_0) - 1/(1-\zeta\xi_0)$ of ξ_0 and η_0 has no singularities on $\mathcal{S}_2$ and so has the same sign on each connected component of $\mathcal{S}_2$ as it does on the portion of $\eta_0 = 0$ that forms part of the boundary of that component. But

$$\delta_+(\xi_0,0) - \frac{1}{1-\zeta\xi_0} = \begin{cases} \frac{(\zeta-\xi_0)(1-\zeta)\xi_0}{(1-\zeta\xi_0)(1+\zeta)(1-\xi_0)^2} & \text{if } \xi_0 \le 0; \\ \\ \frac{(\zeta-\xi_0)(1+\zeta)\xi_0}{(1-\zeta\xi_0)(1-\zeta)(1+\xi_0)^2} & \text{if } \xi_0 \ge 0, \end{cases} \tag{5.200}$$

which is obviously positive, so the condition $\delta < 1/(1-\zeta\xi_0)$ cannot hold if δ is given by $\delta_+(\xi_0,\eta_0)$. The equation $(1-\zeta\xi_0-\delta_+(\xi_0,\eta_0)^{-1})^2 = (1-\zeta^2)\eta_0^2$ has the sole solution

$$\eta_t^{(2)} = \frac{[1+\zeta^2-2\zeta\xi_0 - \sqrt{4(\zeta-\xi_0)^2+(1-\zeta^2)^2}]^2}{4(1-\zeta^2)}, \tag{5.201}$$

and even this is a solution only when $\xi_0 \ge 0$. The only new discontinuities in $(1-\zeta\xi_0-\delta_+(\xi_0,\eta_0)^{-1})^2 - (1-\zeta^2)\eta_0^2$ are the zeros of $\delta_+(\xi_0,\eta_0)$, which are at $\xi_0 = -\frac{1}{2}(1-\zeta)$. To determine the sign of $(1-\zeta\xi_0-\delta_+(\xi_0,\eta_0)^{-1})^2 - (1-\zeta^2)\eta_0^2$ it suffices to examine it along $\eta_0 = 0$ and along the lower part of the boundary of $\mathcal{S}_1 \cap \mathcal{S}_0$, given by $\eta_0 = \sqrt{\eta_{r-}^{(2)}(\xi_0)}$ for $\xi_0 \in \left[-\frac{1}{2}(1-\zeta),\zeta\right)$. Along $\eta_0 = 0$

$$(1-\zeta\xi_0-\delta_+(\xi_0,0)^{-1})^2 - (1-\zeta^2)\eta_0^2 = \begin{cases} \frac{(1-\zeta)^2(\zeta-\xi_0)^2\xi_0^2}{(1+\zeta-2\xi_0)^2} & \text{if } \xi_0 \le 0; \\ \\ \frac{(1+\zeta)^2(\zeta-\xi_0)^2\xi_0^2}{(1-\zeta+2\xi_0)^2} & \text{if } \xi_0 \ge 0, \end{cases} \tag{5.202}$$

which is positive, so the condition $(1-\zeta\xi_0-\delta^{-1})^2 < (1-\zeta^2)\eta_0^2$ cannot hold when δ is given by $\delta_+(\xi_0,\eta_0)$ and either $\xi_0 < 0$, or $\xi_0 \ge 0$ and $\eta_0 \le \sqrt{\eta_t^{(2)}(\xi_0)}$. Along $\eta_0 = \sqrt{\eta_{r-}^{(2)}(\xi_0)}$

$$(1-\zeta\xi_0-\delta_+(\xi_0,\sqrt{\eta_{r-}^{(2)}(\xi_0)})^{-1})^2 - (1-\zeta^2)\eta_0^2 = \frac{(\zeta-\xi_0)^2\left(\sqrt{1-\zeta^2}-\sqrt{1-\zeta^2+4\zeta\xi_0-4\xi_0^2}\right)^2}{(1-2\xi_0+\zeta)(1+2\xi_0-\zeta)}, \tag{5.203}$$

which is positive for $\xi_0 \in \left[-\frac{1}{2}(1-\zeta),\zeta\right)$. Thus, if δ is given by $\delta_+(\xi_0,\eta_0)$, the condition $(1-\zeta\xi_0-\delta^{-1})^2 < (1-\zeta^2)\eta_0^2$ cannot hold anywhere on $\mathcal{S}_2$. Consequently, the condition $1/(1-\zeta\xi_0-\sqrt{1-\zeta^2}\eta_0) > \delta$ cannot hold anywhere on $\mathcal{S}_2$ when δ is given by $\delta_+(\xi_0,\eta_0)$, and so δ cannot be given by $\delta_+(\xi_0,\eta_0)$.

By a similar argument, δ cannot be given by $\delta_-(\xi_0, \eta_0)$, and so (ξ_3, η_3) cannot be a Type 0b point.

■

Theorem 5.29: If more than one point of the set $\mathbb{H}_1 \in \{\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-\}$ is a Type 1 point and more than one point of the $\mathbb{H}_2 \in \{\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-\} - \{\mathbb{H}_1\}$ is a Type 1 point, then no point of $\mathbb{H}_3 \in \{\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-\} - \{\mathbb{H}_1\ \mathbb{H}_2\}$ can be a Type 1 point. □

Proof 5.29: *Suppose more than one point of the set $\mathbb{H}_1$ is of Type 1, where $\mathbb{H}_1 = \pi^{-1}(\{z \in \mathbb{R}^2: [1,0]z = \zeta\})$ (if necessary, an orthogonal transformation O_1 may be applied to bring this about). Then, by Theorem 5.25,*

$$\alpha = \frac{1 - \delta(1 - \zeta^2 - \eta_0^2)}{(\zeta - \xi_0)^2}, \tag{5.204}$$

$$\beta = \delta, \tag{5.205}$$

$$\gamma = \frac{\delta\eta_0}{\zeta - \xi_0}. \tag{5.206}$$

Suppose more than one point of the set $\mathbb{H}_2$ is of Type 1, and there exists a point (ξ_3, η_3) of $\mathbb{H}_3$ which is of Type 1. There are two possibilities:

$\mathbb{H}_2$ is parallel to $\mathbb{H}_1$ and $\mathbb{H}_3$ is not parallel to $\mathbb{H}_1$;

$\mathbb{H}_2$ is not parallel to $\mathbb{H}_1$ and $\mathbb{H}_3$ is parallel to $\mathbb{H}_1$ or to $\mathbb{H}_2$. Without loss of generality, let $\mathbb{H}_3$ be parallel to $\mathbb{H}_1$.

First, suppose $\mathbb{H}_2$ is parallel to $\mathbb{H}_1$. Then, $\mathbb{H}_2 = \pi^{-1}(\{z \in \mathbb{R}^2: [1,0]z = \zeta_2\})$, where $\zeta_2 < \zeta$. Again, by Theorem 5.25,

$$\alpha = \frac{1 - \delta(1 - \zeta_2^2 - \eta_0^2)}{(\zeta_2 - \xi_0)^2}, \tag{5.207}$$

$$\beta = \delta, \tag{5.208}$$

$$\gamma = \frac{\delta\eta_0}{\zeta_2 - \xi_0}. \tag{5.209}$$

As $\zeta_2 \neq \zeta$, the γ-equations (5.206) and (5.209) mean that $\gamma = \eta_0 = 0$. Then the α-equations 5.204 and 5.207 mean that either $\zeta_2 = -\zeta < 0$ and

$$\alpha = \frac{1 - \delta(1 - \zeta^2)}{\zeta^2}, \tag{5.210}$$

$$\beta = \delta, \tag{5.211}$$

$$\gamma = 0, \tag{5.212}$$

$$\xi_0 = 0, \tag{5.213}$$

$$\eta_0 = 0, \tag{5.214}$$

or δ is given by

$$\delta = \frac{\zeta + \zeta_2 - 2\xi_0}{\zeta + \zeta_2 - 2(1 + \zeta\zeta_2)\xi_0 + (\zeta + \zeta_2)\xi_0^2} \tag{5.215}$$

so

$$\alpha = \frac{\zeta + \zeta_2}{\zeta + \zeta_2 - 2(1 + \zeta\zeta_2)\xi_0 + (\zeta + \zeta_2)\xi_0^2} \tag{5.216}$$

$$\beta = \frac{\zeta + \zeta_2 - 2\xi_0}{\zeta + \zeta_2 - 2(1 + \zeta\zeta_2)\xi_0 + (\zeta + \zeta_2)\xi_0^2}, \tag{5.217}$$

$$\gamma = 0 \tag{5.218}$$

$$\eta_0 = 0. \tag{5.219}$$

Since $\mathbb{H}_3$ is not parallel to $\mathbb{H}_1$, $\mathbb{H}_3 = \pi^{-1}(\{z \in \mathbb{R}^2 : [\cos\theta, \sin\theta]z = \zeta_3\})$, for some θ such that $\sin\theta \neq 0$.

Suppose that (ξ_3, η_3) is a Type 1a point. Then, by Theorem 5.21,

$$\alpha(\xi_3 - \xi_0)^2 + \delta(1 - \xi_3^2) = 1, \tag{5.220}$$

$$\delta\xi_3 - \alpha(\xi_3 - \xi_0) = k\cos\theta, \tag{5.221}$$

$$0 = k\sin\theta, \tag{5.222}$$

when equations (5.210) to (5.214), or (5.215) to (5.219), hold. Clearly, $k = 0$ as $\sin\theta \neq 0$.

If equations (5.210) to (5.214) hold, then equation (5.221) means that either $\xi_3 = 0$ or $\delta = 1$. But, if $\xi_3 = 0$, equation (5.220) implies that $\delta = 1$. Hence, $\delta = 1$ and $\mathcal{E}_2 = \mathcal{E}_0$.

If equations (5.215) to (5.219) hold, then equation (5.221) means that either $\xi_0 = 0$ or $\xi_3 = \frac{1}{2}(\zeta + \zeta_2)$. If $\xi_0 = 0$, $\mathcal{E}_2 = \mathcal{E}_0$, and, if $\xi_3 = \frac{1}{2}(\zeta + \zeta_2)$, either $\xi_0 = 0$ or $(\zeta + \zeta_2)^2 = 4\zeta\zeta_2$, which is impossible, as $\zeta_2 \neq \zeta$. Consequently, $\mathcal{E}_2 = \mathcal{E}_0$ again.

Now suppose (ξ_3, η_3) is a Type 1b point. Then equation (5.220) holds and $\xi_3^2 + \eta_3^2 = 1$. If equations (5.210) to (5.214) hold, then equation (5.220) becomes $(\delta - 1)(\zeta^2 - \xi_3^2) = 0$. If $\xi_3 = \zeta$, (ξ_3, η_3) cannot be a Type 1 point, and, as $\zeta_2 = -\zeta$, the same is true if $\xi_3 = -\zeta$. But then $\delta = 1$ and $\mathcal{E}_2 = \mathcal{E}_0$.

If equations (5.210) to (5.214) hold, then equation (5.220) becomes $\xi_0(\xi_3 - \zeta)(\xi_3 - \zeta_2) = 0$ and so $\xi_0 = 0$, as again $\xi_3 \neq \zeta, \zeta_2$. But then $\mathcal{E}_2 = \mathcal{E}_0$.

Now suppose $\mathbb{H}_2$ is not parallel to $\mathbb{H}_1$. Then $\mathbb{H}_2 = \pi^{-1}(\{z \in \mathbb{R}^2 : [\cos\theta, \sin\theta]z = \zeta_2\})$, for some θ such that $\sin\theta \neq 0$. *Then $\eta_2 = (\zeta_2 - \xi_2\cos\theta)/\sin\theta$ for points $(\xi_2, \eta_2) \in \pi(\mathbb{H}_2)$.*

By Theorem 5.21,

$$[(\alpha - \delta)\sin\theta - 2\gamma\cos\theta]\xi_2^2 +$$

$$2[(\gamma\xi_0+\delta\eta_0)\cos\theta-(\alpha\xi_0+\gamma\eta_0)\sin\theta+\gamma\zeta_2]\xi_2+$$
$$(\alpha\xi_0^2+2\gamma\xi_0\eta_0+\delta\eta_0^2+\delta-1)\sin\theta-2\zeta_2(\gamma\xi_0+\delta\eta_0) = 0 \quad (5.223)$$
$$(\delta-\alpha+\gamma\cot\theta)\xi_2+\alpha\xi_0+\gamma(\eta_0-\csc\theta) = k_2(\eta_2)\cos\theta \quad (5.224)$$
$$-\gamma\xi_2+\gamma\xi_0+\delta\eta_0 = k_2(\eta_2)\sin\theta \quad (5.225)$$

for some $k_2(\eta_2)\leq 0$, where equation (5.205) has been used to aid in the simplification. When k_3 is eliminated between equations (5.224) and (5.225),

$$[(\delta-\alpha)\sin\theta+2\gamma\cos\theta]\xi_2+(\alpha\xi_0+\gamma\eta_0)\sin\theta-(\gamma\xi_0+\delta\eta_0)\cos\theta-\gamma\zeta_2=0 \quad (5.226)$$

results. As equations (5.223) to (5.225) are to hold for more than one value of η_2, both the coefficient of η_2 and the constant term in equation (5.226) must vanish. The result of substituting the rest of equations (5.204) to (5.206) into equation (5.223) and the coefficient of η_2 and the constant term in equation (5.226) is

$$[2\eta_0\xi_2(\xi_2-\zeta)(\xi_0-\zeta)\cos\theta+$$
$$(\xi_2-\zeta)(-\xi_2+2\xi_0-\xi_2\xi_0^2+\xi_2\eta_0^2+2\xi_2\zeta\xi_0-\zeta-\eta_0^2\zeta-\zeta\xi_0^2)\sin\theta-$$
$$2(\xi_2-\zeta)\eta_0\zeta_2(-\zeta+\xi_0)]\delta-$$
$$(\xi_2-\zeta)(-\zeta-\xi_2+2\xi_0)\sin\theta = 0, \quad (5.227)$$

$$\eta_0\zeta[-\zeta+\xi_0\cos\theta+(-\xi_0+\zeta^2\xi_0+\eta_0^2\zeta)\sin\theta+\eta_0\zeta_2(-\zeta+\xi-0)]\delta+$$
$$\xi_0\sin\theta = 0, \quad (5.228)$$

$$[-2\eta_0(-\zeta+\xi_0)\cos\theta+(1-\eta_0^2-2\zeta\xi_0+\xi_0^2)\sin\theta]\delta-\sin\theta = 0. \quad (5.229)$$

As $\sin\theta\neq 0$, the coefficient of δ in equation (5.229) cannot vanish and so equation (5.229) can be used to obtain an expression for δ. When this is done, the result can be put into equations (5.204) to (5.206) to obtain

$$\alpha = \frac{(\zeta-\xi_0)\sin\theta+2\eta_0\cos\theta}{(\zeta-\xi_0)[(1-2\zeta\xi_0+\xi_0^2-\eta_0^2)\sin\theta+2(\zeta-\xi_0)\eta_0\cos\theta]}, \quad (5.230)$$
$$\beta=\delta = \frac{\sin\theta}{(1-2\zeta\xi_0+\xi_0^2-\eta_0^2)\sin\theta+2(\zeta-\xi_0)\eta_0\cos\theta}, \quad (5.231)$$
$$\gamma = \frac{\eta_0\sin\theta}{(\zeta-\xi_0)[(1-2\zeta\xi_0+\xi_0^2-\eta_0^2)\sin\theta+2(\zeta-\xi_0)\eta_0\cos\theta]}. \quad (5.232)$$

If the new expression for δ is substituted into equation (5.227) or (5.227),

$$\xi_0^2 \sin^2\theta - 2\xi_0\eta_0 \sin\theta\cos\theta - \eta_0^2 \sin^2\theta + \eta_0(\zeta_2 + \zeta\cos\theta)\sin\theta - \zeta\xi_0 \sin^2\theta = 0 \qquad (5.233)$$

results.

If (ξ_3, η_3) is a Type 1 point, $(\xi_3, \eta_3) \in \pi(\mathbb{H}_3) = \{z \in \mathbb{R}^2 \colon [1,0]z = \zeta_3\}$ for some $\zeta_3 < \zeta$, so $\xi_3 = \zeta_3$. Suppose (ξ_3, η_3) is a Type 1a point. Then, by Theorem 5.21, $\delta\eta_3 - \gamma(\zeta_3 - \xi_0) - \beta(\eta_3 - \eta_0) = 0$, or using equations (5.230) to (5.232), $\eta_0(\zeta - \zeta_3)\sin\theta = 0$, which means that $\eta_0 = 0$, as $\zeta_3 \neq \zeta$, $\sin\theta \neq 0$. But $\eta_0 = 0$ in equation (5.233) yields $(\zeta - \zeta_3)(\zeta - \xi_0)\xi_0 \sin\theta = 0$, which means that $\xi_0 = 0$ as $\zeta \neq \zeta_3, \xi_0$, and this means that $\mathcal{E}_2 = \mathcal{E}_0$.

Now suppose (ξ_3, η_3) is a Type 1b point. Then, $\zeta_3^2 + \eta_3^2 = 1$ and, by Theorem 5.21, $\alpha(\zeta_3 - \xi_0)^2 + 2\gamma(\zeta_3 - \xi_0)(\pm\sqrt{1-\zeta_3^2} - \eta_0)^2 + \beta(\pm\sqrt{1-\zeta_3^2} - \eta_0)^2 = 1$, or, using equations (5.230) to (5.232),

$$\pm\eta_0\sqrt{1-\zeta_3^2}\sin\theta + \eta_0(\zeta - \zeta_3 - 2\xi_0)\cos\theta + (\xi_0^2 - \zeta\xi_0 - \eta_0^2)\sin\theta = 0. \qquad (5.234)$$

When this equation is multiplied by $\sin\theta$ and equation (5.233) is subtracted,

$$[\zeta_3\cos\theta \pm \sqrt{1-\zeta_3^2}\sin\theta - \zeta_2]\eta_0\sin\theta = 0 \qquad (5.235)$$

is obtained. If the quantity in the square brackets here were to vanish, (ξ_3, η_3) would not be a Type 1b point, so $\eta_0 = 0$, again implying that $\xi_0 = 0$ and $\mathcal{E}_2 = \mathcal{E}_0$. ■

Theorem 5.30: If there is a Type 0b point, the set of remaining contact points cannot

1. consist of two Type 1b points on the same hyperplane;

2. consist of three or more Type 1b points distributed over a pair of opposite hyperplanes.

If there are no Type 0b points, the set of contact points cannot

3. consist of three or four Type 1b points belonging to a pair of opposite hyperplanes.

□

Proof 5.30: *If $S = \pi(\partial\mathcal{E}_2 \cap \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2))$ consists entirely of Type ib points, $\pi^{-1}(S) \cap \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$ consists entirely of points in a $(p-2)$-dimensional subspace of $\mathbb{R}^p$, and so the minimum-volume ellipsoid about $\pi^{-1}(S) \cap \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$ cannot be the minimum-volume ellipsoid about $\partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$. This means that at least one of the tangency conditions associated with the Type ib points must be active (that is, at least one of the k's in the tangency conditions must be zero), and so, by Remark 1, either a Type 0b point is governed by equations like those for a Type 0a*

point, or a Type 1b point behaves like a Type 1a point or a Type 0b point. But, by Theorem 5.23, a Type 0b point cannot behave like a Type 0a point, and so there must be a Type 1b point behaving like a Type 1a point or a Type 0b point.

Suppose there is already a Type 0b point. Then, by Theorem 5.24, there must be a Type 1b point behaving like a Type 1a point. If there were just two Type 1b points on a single hyperplane (in addition to the Type 0b point), or three or four Type 1b points distributed across a pair of opposite hyperplanes, the argument of Theorem 5.27 would hold, so there cannot be just two Type 1b points on a single hyperplane, or three or four Type 1b points distributed across a pair of opposite hyperplanes.

Suppose there is not already a Type 0b point. If one of the Type 1b points behaves like a Type 1a point, then, if there are four Type 1b points belonging to a pair of parallel hyperplanes, the argument of the previous paragraph holds. On the other hand, if one of a pair of Type 1b points lying on the same hyperplane behaves like a Type 0b point, then, after a co-ordinate rotation if necessary, the following equations hold:

$$\alpha = \frac{1-\beta(1-\zeta^2)}{(\zeta-\xi_0)^2} \tag{5.236}$$

$$\gamma = \frac{\beta\eta_0}{\zeta-\xi_0}, \tag{5.237}$$

(by Theorem 5.26 — and the fact that the equations for a Type 0b point actually imply those for a Type 1b point)

$$\delta\zeta - \alpha(\zeta-\xi_0) - \gamma(\pm\eta-\eta_0) = k_{1\pm} + k_{2\pm}[\alpha(\zeta-\xi_0)+\gamma(\pm\eta-\eta_0)]. \tag{5.238}$$

$$\pm\delta\eta - \gamma(\zeta-\xi_0) - \beta(\pm\eta-\eta_0) = k_{2\pm}[\gamma(\zeta-\xi_0)+\beta(\pm\eta-\eta_0)], \tag{5.239}$$

(by Theorems 5.20 and 5.21) where $k_{1\pm} \leq 0$ *and* $k_{2+} = 0$ *and* $k_{2-} \leq 0$*, or* $k_{2+} \leq 0$ *and* $k_{2-} = 0$*. Substituting equations (5.236) and (5.237) in equation (5.239) reveals that* $\beta = \delta$*, which would mean the existence of a* $\mathbb{H}T$ *set, a contradiction which proves the Theorem, except for the cases where there are three Type 1b points, as then it is possible that the Type 1b point which behaves like a Type 1a point is opposite the other two Type 1b points.*

When there are three Type 1b points distributed over a pair of parallel hyperplanes, the above argument can be used to show that one of them behaves like a Type 1a point or a Type 0b point. If it behaves like a Type 0b point, one of the others must behave like a Type 1a point. Then the rest of the argument can be adapted to prove the Theorem. ■

Theorem 5.31: There are at most four distinct Type ib points. □

Proof 5.31: *Suppose there are five distinct Type* ib *points. After making a rotation if necessary,*

these can be chosen to be $(\xi_1, \eta_1), (\xi_1, -\eta_1), (\xi_2, \eta_2), (\xi_3, \eta_3)$, *and* (ξ_4, η_4), *where* $\eta_1 > 0$, $\xi_2, \xi_3, \xi_4 \neq \xi_1$ *and* $\xi_i^2 + \eta_i^2 = 1$, $i = 1, 2, 3, 4$, *so*

$$\alpha(\xi_i - \xi_0)^2 + 2\gamma(\xi_i - \xi_0)(\eta_i - \eta_0) + \beta(\eta_i - \eta_0)^2 = 1, \tag{5.240}$$

for $i = 1, \ldots, 4$ *and*

$$\alpha(\xi_1 - \xi_0)^2 + 2\gamma(\xi_1 - \xi_0)(-\eta_1 - \eta_0) + \beta(-\eta_1 - \eta_0)^2 = 1. \tag{5.241}$$

If $\xi_0 = 0$, *equation (5.240) with* $i = 1$ *yields* $\beta = 1/(\eta_1 - \eta_0)^2$, *and putting this in equation (5.241) gives* $4\eta_0\eta_1/(\eta_1 - \eta_0)^2 = 0$, *which means that* $\eta_0 = 0$ *as* $\eta_1 > 0$. *Then* $\beta = 1/\eta_1^2$.
When the above results are put in equation (5.240) with $i = 2$, *the facts that* $\xi_2 \neq \xi_1$ *and* $\eta_i^2 = 1 - \xi_i^2$ *allows the deduction that*

$$\alpha = \frac{2\eta_2(1 - \xi_1^2)\gamma + \xi_1 + \xi_2}{\eta_1^2(\xi_2 - \xi_1)}. \tag{5.242}$$

If this is substituted, along with the results for β, ξ_0 *and* η_0, *in equation (5.240) with* $i = 3$, *it may be deduced that*

$$2\frac{\xi_1 - \xi_3}{\xi_1 - \xi_2}\left(((\eta_2 - \eta_3)\xi_1 - \xi_3\eta_2 + \eta_3\xi_2)\gamma + \frac{\xi_1(\xi_3 - \xi_2)}{1 - \xi_1^2}\right) = 0. \tag{5.243}$$

Now, suppose the coefficient of γ *vanishes in this equation. This means that the term not involving* γ *also vanishes, so either* $\xi_1 = 0$ *or* $\xi_3 = \xi_2$. *The second case means that* $\eta_3 = -\eta_2$, *and then the vanishing of the coefficient of* γ *means that* $\eta_2(\xi_1 - \xi_2) = 0$, *so* $\eta_2 = 0$ *and* $(\xi_3, \eta_3) = (\xi_2, \eta_2)$, *a contradition for distinct points. The first case means that* $\eta_1 = 1$, *and the vanishing of the coefficient of* γ *then means* $\xi_3\eta_2 = \xi_2\eta_3$, *or* $\xi_3 = -\xi_2$, $\eta_3 = -\eta_2$. *Then* $\alpha = 1 - 2\eta_3\gamma/\xi_3$, $\beta = 1$, $\xi_0 = \eta_0 = 0$, *and putting these into equation (5.240) for* $i = 4$ *results in the equation* $\gamma\xi_4(\xi_2\eta_4 - \xi_4\eta_2) = 0$. *But* $\xi_4 \neq 0$ *as* $\xi_1 = 0$ *and* $\xi_2\eta_4 - \xi_4\eta_2 = 0$ *would imply either* $\xi_4 = \xi_2$ *and* $\eta_4 = \eta_2$ *or* $\xi_4 = -\xi_2 = \xi_3$ *and* $\eta_4 = -\eta_2 = \eta_3$, *so* $\gamma = 0$ *and consequently* $\mathcal{E}_2 = \mathcal{E}_0$.
Now suppose the coefficient of γ *in equation (5.243) does not vanish. Then*

$$\gamma = \frac{(\xi_2 - \xi_3)\xi_1}{\eta_1^2(\xi_1(\eta_2 - \eta_3) - \xi_3\eta_2 + \eta_4\xi_3)}, \tag{5.244}$$

and putting this and equation (5.242) in equation (5.240) with $i = 4$ *yields*

$$\frac{(\xi_4 - \xi_1)(\xi_4\eta_3 + \eta_4\xi_3 - \xi_4\eta_2 - \eta_3\xi_2 + \xi_3\eta_2 - \eta_4\xi_3)\xi_1}{(1 - \xi_1^2)((\eta_2 - \eta_3)\xi_1 - \xi_3\eta_2 + \eta_3\xi_2)} = 0. \tag{5.245}$$

As $\xi_4 \neq \xi_1$, *the numerator of the left-hand side of this equation only disappears when* $|\xi_4| = |\xi_3| = |\xi_2| = 1$, *which is not possible.*

Thus, if $\xi_0 = \xi_1$ a contradiction results.

Suppose now $\xi_0 \neq \xi_1$. Then, the simultaneous solution of equation (5.240) with $i = 1$ and equation (5.241) yields

$$\alpha = \frac{1 - \beta(1 - \xi_1^2 - \eta_0^2)}{(\xi_1 - \xi_0)^2}, \tag{5.246}$$

$$\gamma = \frac{\beta\eta_0}{\xi_1 - \xi_0}. \tag{5.247}$$

Substituting this in equation (5.240) with $i = 2$ yields

$$\frac{\xi_1 - \xi_2}{(\xi_1 - \xi_0)^2} \times \left[((\xi_1 + \xi_2)\xi_0^2 + 2\eta_2\xi_0\eta_0 + (\xi_1 - \xi_2)\eta_0^2 - 2(1 + \xi_1\xi_2)\xi_0 - 2\xi_1\xi_2\eta_0 + \xi_1 + \xi_2)\,\beta - \xi_1 - \xi_2 + 2\xi_0\right] = 0. \tag{5.248}$$

There are similar expressions with ξ_2, η_2 replaced by ξ_3, η_3 and ξ_4, η_4, and, by symmetry, if the coefficient of β does not vanish in all three such expressions, it may be assumed without loss of generality that it does not vanish in equation (5.248). But, if the coefficient of β vanishes in all three expressions, then $\xi_1 + \xi_2 - 2\xi_0 = \xi_1 + \xi_3 - 2\xi_0 = \xi_1 + \xi_4 - 2\xi_0 = 0$, which would mean that $\xi_4 = \xi_3 = \xi_2$, which not possible.

Thus, it can be assumed that equation (5.248) may be solved for β. If this is done and the result is substituted into equation (5.240) with $i = 3$ or $i = 4$, the following equation will be seen to hold:

$$(\xi_i - \xi_2)\eta_0^2 + (2(\eta_2 - \eta_i)\xi_0 + (\eta_i - \eta_2)\xi_1 + \eta_i\xi_2 - \xi_i\eta_2)\eta_0 + \xi_0((\xi_i - \xi_2)(\xi_1 - \xi_0) = 0. \tag{5.249}$$

The result of eliminating η_0^2 between this equation with $i = 3$ and $i = 4$ is

$$\eta_0(\xi_1 + \xi_2 - 2\xi_0)(-\xi_4\eta_2 - \eta_3\xi_2 + \eta_4\xi_2 - \eta_4\xi_3 + \xi_4\eta_3 + \xi_3\eta_2) = 0. \tag{5.250}$$

If the solution $\eta_0 = 0$ is chosen here, it will be seen that $\xi_0 = 0$ as well, by substituting $\eta_0 = 0$ back into the previous equation, and then $\mathcal{E}_2 = \mathcal{E}_0$.

If the solution $\xi_0 = \frac{1}{2}(\xi_1 + \xi_2)$ is substituted back, it will be seen that $(\xi_3 - \xi_2)(4\eta_0^2 - 4\eta_2\eta_0 + \xi_1^2 - \xi_2^2) = (\xi_4 - \xi_2)(4\eta_0^2 - 4\eta_2\eta_0 + \xi_1^2 - \xi_2^2) = 0$. As $\xi_4 = \xi_3 = \xi_2$ is impossible, $(4\eta_0^2 - 4\eta_2\eta_0 + \xi_1^2 - \xi_2^2) = 0$, which means $\eta_0 = (\eta_2 \pm \eta_1)/2$. Both of these solutions for η_0, when substituted into the current expressions for β, imply that $\beta = 0$, a contradiction.

Thus, there are at most 4 Type ib points. ■

Theorem 5.32: If there is a Type 0b point, there are at most three distinct Type ib points. □

Proof 5.32: *Suppose there is a Type 0b point and at least four Type ib points, including the Type 0b point. By Theorem 5.24, there is exactly one Type 0b point. In a similar fashion to the*

proof of Theorem 5.31, the Type 0b point can be set to (ξ_1, η_1)*, and the other three Type* ib *points to* $(\xi_1, -\eta_1)$*,* (ξ_2, η_2) *and* (ξ_3, η_3)*. Then, an equation like equation (5.240) will hold for* $i = 1, 2, 3$ *and equation (5.241) will also hold. The equation*

$$\delta \begin{bmatrix} \xi_1 \\ \eta_1 \end{bmatrix} - \begin{bmatrix} \alpha(\xi_1 - \xi_0) + \gamma(\eta_1 - \eta_0) \\ \gamma(\xi_1 - \xi_0) + \beta(\eta_1 - \eta_0) \end{bmatrix} = k \begin{bmatrix} \alpha(\xi_1 - \xi_0) + \gamma(\eta_1 - \eta_0) \\ \gamma(\xi_1 - \xi_0) + \beta(\eta_1 - \eta_0) \end{bmatrix} \tag{5.251}$$

will hold for some $k \leq 0$ *(replacing equation (5.240) for* $i = 3$ *in Theorem 5.31).*

The method of Theorem 5.31 can be followed up to points before equation (5.240) for $i = 3$ *was deployed. Hence,*

either $\xi_0 = \xi_1$ *In this case,* $\eta_0 = \xi_0 = \xi_1 = 0$*,* $\eta_1 = 1$*,* $\xi_3 = -\xi_2$*,* $\eta_3 = -\eta_2$*,* $\beta = 1$ *and* $\alpha = 1 - 2\eta_2\gamma/\xi_3$*. This means that equation (5.251) becomes*

$$\begin{bmatrix} -\gamma \\ \delta - \beta \end{bmatrix} = k \begin{bmatrix} \gamma \\ \beta \end{bmatrix} \tag{5.252}$$

which has the solutions $k = -1$*,* $\delta = 0$*, which is impossible, and* $\gamma = 0$*,* $k = (\delta - \beta)/\beta$*. The latter solution implies* $\mathcal{E}_2 = \mathcal{E}_0$*, a contradiction;*

or $\xi_0 \neq \xi_1$ *here equations (5.246) and (5.247) hold. These can be substituted into equation (5.251) to yield*

$$\begin{bmatrix} \frac{\delta\xi_1(\xi_1 - \xi_0) - 1 + \beta(1 - \xi_1^2 - \eta_1\eta_0)}{\xi_1 - \xi_0} \\ (\delta - \beta)\eta_1 \end{bmatrix} = k \begin{bmatrix} \frac{1 - \beta(1 - \xi_1^2 - \eta_1\eta_0)}{\xi_1 - \xi_0} \\ \beta\eta_1 \end{bmatrix}, \tag{5.253}$$

which has the solution $k = (\delta - \beta)/\beta$*,* $\beta = 1/(1 - \xi_1\xi_0 - \eta_1\eta_0)$*. Now,* $(\xi_2 - \xi_1)$ *times equation (5.240) with* $i = 3$ *minus* $(\xi_3 - \xi_1)$ *times equation (5.240) with* $i = 2$ *yields* $\eta_0(\xi_1 - \xi_0)(-\xi_1\eta_2 + \xi_2\eta_1 + \xi_1\eta_3 - \xi_2\eta_3 - \xi_3\eta_1 + \xi_3\eta_2) = 0$*. If* $\eta_0 = 0$*, equation (5.240) with* $i = 2$ *becomes* $\xi_0(\xi_2 - \xi_1)(\xi_1 - \xi_0) = 0$*. But, if* $\xi_0 = 0$*,* $\mathcal{E}_2 = \mathcal{E}_0$*, and* $(\xi_1 - \xi_0)$ *and* $(-\xi_1\eta_2 + \xi_2\eta_1 + \xi_1\eta_3 - \xi_2\eta_3 - \xi_3\eta_1 + \xi_3\eta_2)$ *cannot vanish without violating the current assumptions.*

Thus, the assumption that there are at least four Type ib *points leads to a contradiction, establishing the Theorem.* ■

There are two more Theorems like this, which can be proven in a similar way:

Theorem 5.33: If there is a HIT set, then there can be at most two Type ib points. □

Theorem 5.34: If there is a HIT set and a Type 1a set, then there can be no Type ib points. □

The eventual aim is to find a set of polynomial equations in the six variables $\{\alpha, \beta, \gamma, \delta, \xi_0, \eta_0\}$ and reducing this set as far as possible to find a solution of the problem of minimising the volume of the corresponding ellipsoid.

As a first step in this direction, it is necessary to rewrite the conditions satisfied by points of Types 0b and 1a.

Theorem 5.35: Suppose there exists a point of Type 0b. Then

$$(\mathrm{Trace}\bar{Q}^{-1})^2(\bar{c}^{\mathrm{T}}\bar{c})^3(\bar{c}^{\mathrm{T}}\bar{c}-4)-$$
$$16[(\bar{c}^{\mathrm{T}}\bar{Q}^{-1}\bar{c})^2-(\mathrm{Trace}\bar{Q}^{-1})(\bar{c}^{\mathrm{T}}\bar{Q}^{-1}\bar{c})(\bar{c}^{\mathrm{T}}\bar{c})+(\det\bar{Q}^{-1})(\bar{c}^{\mathrm{T}}\bar{c})^2]=0. \tag{5.254}$$

□

Proof 5.35: *Suppose the Type 0b point is $\bar{x}=(\xi_1,\eta_1)$, where, of course, $\xi_1^2+\eta_1^2=1$. An orthogonal transformation O_1 may be applied to make $\gamma=0$. This will be used in*

$$(\bar{x}_1-\bar{c})^{\mathrm{T}}\bar{Q}^{-1}(\bar{x}_1-\bar{c})=1, \tag{5.255}$$

and $\delta\bar{x}_1-\bar{Q}^{-1}(\bar{x}_1-\bar{c})=k\bar{Q}^{-1}(\bar{x}_1-\bar{c})$ for some $k\leq 0$. This last expression is equivalent to

$$(\delta\mathrm{I}_2-k'\bar{Q}^{-1})\bar{x}_1=k'\bar{Q}^{-1}\bar{c},\ \ k'\leq 1. \tag{5.256}$$

In addition,

$$\bar{x}^{\perp\mathrm{T}}(\delta\mathrm{I}_2-\bar{Q}^{-1})\bar{x}^{\perp}\geq 0,\qquad \text{if } k'<1;$$
$$\delta\mathrm{I}_2-\bar{Q}^{-1}\geq 0,\qquad \text{if } k'=1, \tag{5.257}$$

where $\bar{x}^{\perp}=(-\eta_1,\xi_1)$.

One of the following must hold: $\delta\mathrm{I}_2-k'\bar{Q}^{-1}=0$; $\delta\mathrm{I}_2-k'\bar{Q}^{-1}$ is degenerate but nonzero; $\delta\mathrm{I}_2-k'\bar{Q}^{-1}$ is non-degenerate. It will be shown that the first two possibilities cannot occur and that the third leads to the expression of the statement of the Theorem.

$\delta\mathbf{I}_2-k'\bar{Q}^{-1}=0$: *Here equation (5.256) yields $\delta=k'\alpha$ and $\beta=\alpha$, and $\xi_0=\eta_0=0$ as $\bar{Q}$ is positive definite. A consequence of $\delta=k'\alpha$ is that $k'\in(0,1]$. Equation (5.255) means that $\alpha=1$, so $\det Q^{-1}=\alpha\beta\delta^{p-2}=k'^{p-2}\leq 1$, a contradiction.*

$\delta\mathbf{I}_2-k'\bar{Q}^{-1}\neq 0$, $\det(\delta\mathbf{I}_2-k'\bar{Q}^{-1})=0$: *Here, $\beta\neq\alpha$ and either $\delta=k'\alpha$ or $\delta=k'\beta$. Without loss of generality, $\delta=k'\alpha$ will be chosen, which requires that $k'\in(0,1]$. Then equation (5.256) means that $\xi_0=0$ and $\eta_1=-\beta\eta_0/(\alpha-\beta)$. Equation (5.255) becomes*

$$\alpha+\frac{\alpha\beta}{\alpha-\beta}\eta_0^2=1 \tag{5.258}$$

so

$$\frac{(\alpha-1)(\beta-\alpha)}{\alpha\beta} = \eta_0 \in [0,1), \tag{5.259}$$

which implies either $\beta < \alpha < 1$ *or* $\beta > \alpha > 1$. *The first of these implies that* $\det Q^{-1} = k'^{p-2}\alpha^{p-1}\beta < 1$, *which is a contradiction.*

Thus, $\beta > \alpha > 1$. *But*

$$\frac{\beta(\alpha-1)}{\alpha(\beta-\alpha)} = \eta_1^2 \in [0,1), \tag{5.260}$$

which implies $\beta > \alpha^2$.

Finally, $\bar{x}_1^{\perp\mathrm{T}}(\delta \mathrm{I}_2 - \bar{Q}^{-1})\bar{x}_1 = (\delta-\alpha)\eta_1^2 + (\delta-\beta)\xi_1^2 = (k'-1)\alpha\eta_1^2 + (k'\alpha-\beta)(1-\eta_1^2) = k'\alpha - \beta + (\beta-\alpha)\eta_1^2 = k'\alpha - \beta/\alpha \geq 0$ *which results in the contradiction* $1 \geq k' \geq \beta/\alpha^2 > 1$. *Obviously, setting* $\delta = k'\beta$ *will lead to a similar contradiction.*

$\det(\delta \mathbf{I}_2 - k'\bar{Q}^{-1}) \neq 0$: *Now equation (5.256) can be solved for* $\bar{x}_1$*:*

$$\bar{x}_1 = -k'(\delta \mathrm{I}_2 - k'\bar{Q}^{-1})^{-1}\bar{Q}^{-1}\bar{c}, \tag{5.261}$$

and

$$\bar{x}_1 - \bar{c} = -\delta(\delta \mathrm{I}_2 - k'\bar{Q}^{-1})^{-1}\bar{c} \tag{5.262}$$

also holds.

Therefore, $\bar{x}^{\mathrm{T}}\bar{x} = 1$ *becomes*

$$k'^2\left[\frac{\alpha^2\xi_0^2}{(-\delta+k'\alpha)^2} + \frac{\beta^2\eta_0^2}{(-\delta+k'\beta)^2}\right] = 1 \tag{5.263}$$

and equation(5.255) becomes

$$\delta^2\left[\frac{\xi_0^2}{(-\delta+k'\alpha)^2} + \frac{\eta_0^2}{(-\delta+k'\beta)^2}\right] = 1. \tag{5.264}$$

If $\beta = \alpha$, *then the above equations imply that* $\delta = \pm k'\alpha$. *But* $\delta = k'\alpha$, $\beta = \alpha$ *reduces to the case* $\delta \mathrm{I}_2 - k'\bar{Q}^{-1} = 0$, *which has already been dealt with, and substituting* $\delta = -k'\alpha$ *in equation (5.256) yields* $\bar{x}_1 = \frac{1}{2}\bar{c}$, *which would mean* ${\bar{x}_1}^{\mathrm{T}}\bar{x}_1 = \frac{1}{4}\bar{c}^{\mathrm{T}}\bar{c} < \frac{1}{4}$, *contradiction.*

Thus $\beta \neq \alpha$, *which means that equations (5.263) and (5.264) can be solved for* ξ_0^2 *and* η_0^2*:*

$$\xi_0^2 = \frac{(-\delta+k'\alpha)^2(k'^2\beta^2-\delta^2)}{k'^2\delta^2(\beta^2-\alpha^2)}, \tag{5.265}$$

$$\eta_0^2 = -\frac{(-\delta+k'\beta)^2(k'^2\alpha^2-\delta^2)}{k'^2\delta^2(\beta^2-\alpha^2)}. \tag{5.266}$$

These two equations mean that

$$\xi_0^2 + \eta_0^2 = 2\frac{(k'\alpha - \delta)(\delta - k'\beta)}{k'\delta(\alpha + \beta)} \tag{5.267}$$

which is equivalent to a quadratic in k' which can be solved for k'^2:

$$k'^2 = \delta\frac{(\alpha + \beta)(2 - \xi_0^2 - \eta_0^2)k' - 2\delta}{2\alpha\beta}. \tag{5.268}$$

This equation can then be substituted into the quadratic factors $k'^2\beta^2 - 2\beta\delta k' + \delta^2$ and $\alpha^2 k'^2 - \delta^2$ of the numerator of the right-hand side of equation (5.266) to give two quantities linear in k'. The product of these quantities will again be quadratic in k', and equation (5.268) can then be substituted into this product and the denominator of the right-hand side of equation (5.266). The result of doing this is an equation for η_0^2 the right-hand side of which has a numerator and denominator which are both linear in k'. This equation can be solved for k' to yield:

$$k' = \frac{2\delta[4\beta\xi_0^2 + 4\alpha\eta_0^2 - (\alpha + \beta)(\xi_0^2 + \eta_0^2)^2]}{8\beta\xi_0^2 + 8\alpha\eta_0^2 - 2(\alpha + \beta)(\xi_0^2 + \eta_0^2)((\alpha + 3\beta)\xi_0^2 + (3\alpha + \beta)\eta_0^2) + (\alpha + \beta)^2(\xi_0^2 + \eta_0^2)^3}. \tag{5.269}$$

But now this can be put back into equation (5.267) to obtain

$$(\alpha + \beta)^2(\xi_0^2 + \eta_0^2)^3[(\xi_0^2 + \eta_0^2) - 4] + 16(\alpha - \beta)^2\xi_0^2\eta_0^2 = 0. \tag{5.270}$$

To finish, it is now necessary to express this equation in terms of quantities which are unaffected by rotations O_1 leaving vectors in the subspace orthogonal to e_1 and e_2 unchanged. The quantity $16(\alpha-\beta)^2\xi_0^2\eta_0^2$ can be so expressed by eliminating $\xi_0^2\eta_0^2$ from the equations giving the expansions of $(\bar{c}^{\mathrm{T}}\bar{c})^2$, $(\bar{c}^{\mathrm{T}}\bar{c})(\bar{c}^{\mathrm{T}}\bar{Q}^{-1}\bar{c})$ and $(\bar{c}^{\mathrm{T}}\bar{Q}^{-1}\bar{c})^2$. This procedure supplies $(\alpha-\beta)^2\xi_0^2\eta_0^2 = -(\bar{c}^{\mathrm{T}}\bar{Q}^{-1}\bar{c})^2 + (\mathrm{Trace}\bar{Q}^{-1})(\bar{c}^{\mathrm{T}}\bar{c})(\bar{c}^{\mathrm{T}}\bar{Q}^{-1}\bar{c}) - (\det\bar{Q}^{-1})(\bar{c}^{\mathrm{T}}\bar{c})^2$ and then equation (5.254) can easily be derived from (5.270).

■

Theorem 5.36: Suppose there exists a point of Type 1a belonging to $\mathbb{H} := \pi^{-1}\{z \in \mathbb{R}^2 : n^{\mathrm{T}}z = \zeta\} \in \{\mathbb{H}_1^+, \mathbb{H}_1^-, \mathbb{H}_2^+, \mathbb{H}_2^-\}$, for $n \in \mathbb{R}^2$ such that $n^{\mathrm{T}}n = 1$. Then, either

$$\bar{Q}^{-1} = \frac{1 - \delta(1 - \zeta^2 - \eta_0^2)}{(\zeta - \xi_0)^2}nn^{\mathrm{T}} + \frac{\delta\eta_0}{\zeta - \xi_0}(n^{\perp}n^{\mathrm{T}} + nn^{\perp\mathrm{T}}) + \delta n^{\perp}n^{\perp\mathrm{T}}, \tag{5.271}$$

$$\bar{c} = \xi_0 n + \eta_0 n^{\perp}, \tag{5.272}$$

for some $\delta \in (0, 1/(1-\zeta^2))$ and $\xi_0 n + \eta_0 n^\perp \in \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, or

$$\bar{Q}^{-1} = \frac{(1-\zeta^2)\delta^2 - [(1-\zeta^2-\eta_0^2)\beta + 2\eta_0(\zeta-\xi_0)\gamma + 1]\delta + \beta + \gamma^2(\zeta-\xi_0)^2}{(\beta-\delta)(\zeta-\xi_0)^2} nn^{\mathrm{T}} + \frac{\delta\eta_0}{\zeta-\xi_0}(n^\perp n^{\mathrm{T}} + nn^{\perp\mathrm{T}}) + \delta n^\perp n^{\perp\mathrm{T}}, \tag{5.273}$$

$$\bar{c} = \xi_0 n + \eta_0 n^\perp, \tag{5.274}$$

for some $\delta \in (0,\infty)$, $\beta \in (0,\infty)$, $\gamma \in (-\sqrt{\alpha\beta}, \sqrt{\alpha\beta})$ and $\xi_0 n + \eta_0 n^\perp \in \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, provided that $\alpha = n^{\mathrm{T}}\bar{Q}^{-1}n$ is positive. In both the alternatives, $n^\perp = [-n_2, n_1]$. □

Proof 5.36: *Let a rotation O_1 be applied that takes n into $[1,0]^{\mathrm{T}}$. Let the transformed Type 1a point be (ζ,η). Then the coincidence equation $\alpha(\zeta-\xi_0)^2 + 2\gamma(\zeta-\xi_0)(\eta-\eta_0) + \beta(\eta-\eta_0)^2 + \delta(1-\zeta^2-\eta^2) = 1$ and the tangent $\delta\eta - \gamma(\zeta-\xi_0) - \beta(\eta-\eta_0) = 0$ both hold. If $\beta \neq \delta$, the tangent equation implies*

$$\eta = \frac{\beta\eta_0 - \gamma(\zeta-\xi_0)}{\delta-\beta}, \tag{5.275}$$

and when this is put into the coincidence equation, equation (5.273) is derived by solving for α and applying the inverse of O_1.

On the other hand, if $\beta = \delta$, the tangent condition implies that $\gamma = \delta\eta_0/(\zeta-\xi_0)$. If this and $\beta = \delta$ are put into the coincidence equation, equation (5.271) can be derived.

The remainder of the Theorem is trivial. ■

It will, of course, be noted that the first alternative of the preceding Theorem implies that more than one point of $\mathbb{H}$ is a Type 1a point.

A similar expression, not involving ξ or η, resulting from the existence a Type 1b point (ξ,η) on a hyperplane defined by $n^{\mathrm{T}}z = \zeta$ can be derived, but it is sufficiently complicated that it is better to add a variable (corresponding to η) to the system of equations and supplement the system by $\eta^2 = 1-\zeta^2$.

Now some theorems dealing with the minimum number of contact points will be derived.

Definition 5.6: *Let the distance $\rho(S,T) = \inf_{(x,y)\in S\times T}\sqrt{(x-y)^{\mathrm{T}}(x-y)}$. For a point y, this notation will be slightly abused by writing $\rho(y,T)$ for $\rho(\{y\},T)$.* □

Lemma 5.37: Suppose $\bar{x}_i, i = 1,\ldots,m$ are contact points and suppose that if $\bar{x} \in \pi(\partial[\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2]) - \{\bar{x}_1,\ldots,\bar{x}_m\}$, then $(\bar{x}-\bar{c})^{\mathrm{T}}\bar{Q}^{-1}(\bar{x}-\bar{c}) + \delta(1-\bar{x}^{\mathrm{T}}\bar{x}) < 1$. Obviously, the appropriate Jacobian conditions also hold for $\bar{x}_1,\ldots,\bar{x}_m$.

If there exist an interval $[0,k_0] \neq \emptyset$ and continuous functions $\alpha'(k), \beta'(k), \gamma'(k), \delta'(k), \xi_0'(k)$ and $\eta_0'(k)$ such that $\bar{x}_i, i = 1,\ldots,n$ remain points of $\pi(\mathcal{E}(c'(k),Q'(k)) \cap \partial(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2))$ for $k \in [0,k_0]$,

where

$$c'(k) = \begin{bmatrix} \bar{c}'(k) \\ 0 \end{bmatrix} \in \mathbb{R}^p,\ Q'^{-1}(k) = \begin{bmatrix} \bar{Q}'^{-1}(k) & 0 \\ 0 & \mathrm{I}_{p-2} \end{bmatrix} \tag{5.276}$$

for

$$\bar{c}'(k) = \begin{bmatrix} \xi_0'(k) \\ \eta_0'(k) \end{bmatrix},\ \bar{Q}'^{-1}(k) = \begin{bmatrix} \alpha'(k) & \gamma'(k) \\ \gamma'(k) & \beta'(k) \end{bmatrix}, \tag{5.277}$$

and the appropriate Jacobian conditions continue to hold with $\bar{Q}$, $\bar{c}$ and δ replaced with $\bar{Q}'(k)$, $\bar{c}'(k)$ and $\delta'(k)$, then there exists $k' \in (0, k_0]$ such that $(\bar{x}-\bar{c}'(k))^{\mathrm{T}}\bar{Q}'^{-1}(k)(\bar{x}-\bar{c}'(k))+\delta'(k)(1-\bar{x}^{\mathrm{T}}\bar{x}) < 1$ for all $\bar{x} \in \pi(\partial[\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2]) - \{\bar{x}_1, \ldots, \bar{x}_n\}$ and all $k \in [0, k']$ and the $\bar{x}_i$ remain contact points for $\mathcal{E}(c'(k), Q'(k))$. □

Proof 5.37: *Since $\mathcal{E}_0$ and $\mathcal{E}_2$ are ellipsoids, the $\mathbb{H}_{1,2}^{\pm}$ are hyperplanes and x_i is an isolated contact point, it can be verified that there exists an open neighbourhood V_i of $\bar{x}_i$ such that for all paths $y(\ell)$ in $\pi^{-1}(V_i) \cap \mathcal{E}_2$ for which $y(0) \in \pi^{-1}(\bar{x}_i)$ and $\pi(y(\ell) - y(0))^{\mathrm{T}}\pi(y(\ell) - y(0))$ is a strictly increasing function of ℓ, $\rho(y(\ell), \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$ is a strictly increasing function of ℓ. Let $U_i \subset V_i$ be a closed neighbourhood of x_i, with $x_i \in U_i{}^{\circ}$. It can also be verified that, because of the continuity of the functions $\alpha'(k), \beta'(k), \gamma'(k), \delta'(k), \xi_0'(k)$ and $\eta_0'(k)$, there exists $k_i > 0$ such that for each $k \in [0, k_i]$ the set of paths $y(\ell)$ in $\pi^{-1}(V_i) \cap \mathcal{E}(\bar{c}'(k), \bar{Q}'(k))$ for which $y(0) \in \pi^{-1}(\bar{x}_i)$ and $\pi(y(\ell)-y(0))^{\mathrm{T}}\pi(y(\ell)-y(0))$ is a strictly increasing function of ℓ, is such that $\rho(y(\ell), \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)$ is a strictly increasing function of ℓ.*

Thus, for each $i = 1, \ldots, n$, there exists $k_i \in [0, k_0]$ and an open neighbourhood $U_i{}^{\circ}$ of $\bar{x}_i$ such that for all $\bar{x} \in (U_i{}^{\circ} \cap \pi(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2)) - \{\bar{x}_i\}$, $(\bar{x} - \bar{c}'(k))^{\mathrm{T}}\bar{Q}'^{-1}(k)(\bar{x} - \bar{c}'(k)) + \delta'(k)(1 - \bar{x}^{\mathrm{T}}\bar{x}) < 1$. Let $k'' = \min_{1 \le i \le n}\{k_i\}$. Then, for all $k \in [0, k'']$ and all $\bar{x} \in (\pi(\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2) \cap \bigcup_{1 \le i \le n} U_i{}^{\circ}) - \{\bar{x}_1, \ldots, \bar{x}_n\}$, $(\bar{x} - \bar{c}'(k))^{\mathrm{T}}\bar{Q}'^{-1}(k)(\bar{x} - \bar{c}'(k)) + \delta'(k)(1 - \bar{x}^{\mathrm{T}}\bar{x}) < 1$.

The sets $\mathcal{F} = \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2 - \pi^{-1}(\bigcup_{1 \le i \le n} U_i{}^{\circ})$ and $\partial\mathcal{E}_2$ are compact and so the distance $\rho(\mathcal{F}, \partial\mathcal{E}_2)$ will be positive, 2ϵ, say.

As the functions $\alpha'(k)$, $\beta'(k)$, $\gamma'(k)$, $\delta'(k)$, $\xi_0'(k)$ and $\eta_0'(k)$ are continuous, the distance $\rho(\mathcal{F}, \partial\mathcal{E}(c'(k), Q'(k))) > \epsilon$ for all $k \in [0, k''']$ for some $k''' > 0$. Then the k' of the statement of the Lemma will be $\min\{k'', k'''\}$.

■

Theorem 5.38: There are at least three contact points. □

Proof 5.38: *Clearly, there are at least two contact points. Suppose there are exactly two contact points, $\bar{x}_i = [\xi_i, \eta_i]^{\mathrm{T}}, i = 1, 2$, $\bar{x}_2 \neq \bar{x}_1$. An orthogonal rotation O_1 may be applied to make $\xi_2 = \xi_1$.*

Then $\eta_2 \neq \eta_1$ *and*

$$\alpha(\xi_1 - \xi_0)^2 + 2\gamma(\xi_1 - \xi_0)(\eta_1 - \eta_0) + \beta(\eta_1 - \eta_0)^2 + \delta(1 - \xi_1^2 - \eta_1^2) = 1, \tag{5.278}$$

$$\alpha(\xi_1 - \xi_0)^2 + 2\gamma(\xi_1 - \xi_0)(\eta_2 - \eta_0) + \beta(\eta_2 - \eta_0)^2 + \delta(1 - \xi_1^2 - \eta_2^2) = 1, \tag{5.279}$$

The difference of these two equations implies that

$$\eta_0 = \frac{(\eta_1 + \eta_2)(\beta - \delta) + 2\gamma(\xi_1 - \xi_0)}{2\beta} \tag{5.280}$$

as $\eta_2 \neq \eta_1$ *and* $\beta > 0$.

Substituting equation (5.280) back into equation (5.278) produces a quadratic in ξ_0*, which has a real solution if*

$$\bar{\Delta} = [(\eta_2 - \eta_1)\beta + (\eta_1 + \eta_2)\delta]^2 - 4\beta[1 - \delta(1 - \xi_1^2 - \eta_2^2)] \geq 0. \tag{5.281}$$

(The left-hand side of this inequality is symmetric in η_1 *and* η_2*, despite its initial appearance.)*

As there must be a real solution for ξ_0*,* β *and* δ *can only be varied in such a way that relation (5.281) continues to hold. If the inequality holds in relation (5.281), this is no problem, as then it will only be a question of making sure that* β *and* δ *remain in a certain neighbourhood. However, if equality holds in relation (5.281),* (β, δ) *can only vary in certain directions.*

As $\bar{x}_1$ *and* $\bar{x}_2$ *are contact points, Jacobian conditions must hold. For* $\bar{x}_1$*, these are one of the following sets of conditions:*

1. $\mathrm{I} - \delta^{-1}\bar{Q}^{-1} \geq 0$;

2. (a) $n_i^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n_i^{\perp} \geq 0, i = 1, 2$;

   (b) $n_i^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n_i^{\perp} \geq 0$, $i = 1$ *or* 2, $w^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})w^{\perp} \geq 0$, *where* $w = \bar{Q}^{-1}(\bar{x}_1 - \bar{c})$ *and* $n_1^{\perp}$ *is one of the two (in* $\mathbb{R}^2$*) unit vectors perpendicular to* n_1;

3. (a) $n_i^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n_i^{\perp} \geq 0$, $i = 1$ *or 2*;

   (b) $w^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})w^{\perp} \geq 0$;

4. *no conditions.*

As above, if the relation "$\geq$*" can be replaced by "*$>$*" throughout the set of conditions, the set of conditions will be satisfied if* $(\alpha, \beta, \gamma, \delta)$ *remains within a certain neighbourhood of its original*

value. The consequence of this is that $\det Q^{-1} = \delta^{p-2}(\alpha\beta - \gamma^2)$ *can be increased if "*$\geq$*" can be replaced by "*$>$*" in the set of conditions connected with the contact point* $\bar{x}_1$*, in the set of conditions connected with the contact point* $\bar{x}_2$ *and in relation (5.281).*

Suppose this replacement cannot be made everywhere. It can be shown that combining the conditions associated with $\bar{x}_1$ *with those associated with* $\bar{x}_2$ *results in one of the sets of conditions:*

1. $\mathrm{I} - \delta^{-1}\bar{Q}^{-1} \geq 0$;

2. *(a)* $n_i^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n_i^{\perp} \geq 0, i = 1, 2$;

   (b) $n_i^{\perp\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})n_i^{\perp} \geq 0$, $i = 1$ *or* 2, $w_j^{\perp\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})w_j^{\perp} \geq 0$, *where* $w_j = \bar{Q}^{-1}(\bar{x}_j - \bar{c})$, $j = 1$ *or* 2;

   (c) $w_i^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})w_i^{\perp} \geq 0, i = 1, 2$;

3. *(a)* $n_i^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})n_i^{\perp} \geq 0$, $i = 1$ *or* 2;

   (b) $w^{\perp\mathrm{T}}(\mathrm{I} - \delta^{-1}\bar{Q}^{-1})w^{\perp} \geq 0$;

4. *no conditions.*

Suppose $\mathrm{I} - \delta^{-1}\bar{Q}^{-1} \geq 0$ *holds, but* $\mathrm{I} - \delta^{-1}\bar{Q}^{-1} > 0$ *does not. Then,* $\delta = \alpha$ *or* $\delta = \beta$*, and both of these imply that* $\gamma = 0$*, or* $(\delta - \alpha)(\delta - \beta) - \gamma^2 = 0$*. In any case,* $\delta \geq \alpha, \beta$.

Suppose first that $\delta = \alpha$ *or* $\delta = \beta$*. If* $\bar{\Delta}$ *is not equal to zero,* $\alpha'(k)$, $\beta'(k)$, $\gamma'(k)$, $\delta'(k)$ *can be set to* $\alpha'(k) = \beta'(k) = \delta'(k) = \beta + k$ *and* $\gamma'(k) = 0$*. Then* $\det Q'(k)^{-1} = (\beta + k)^p$ *will increase in an interval such that* $\bar{\Delta}'(k) = [(\eta_2 - \eta_1)\beta'(k) + (\eta_1 + \eta_2)\delta'(k)]^2 - 4\beta'(k)[1 - \delta'(k)(1 - \xi_1^2 - \eta_2^2)] > 0$*. If* $\bar{\Delta} = 0$, $\alpha'(k)$, $\beta'(k)$ *and* $\gamma'(k)$ *can be set to* $\alpha'(k) = \alpha$, $\beta'(k) = \beta + k$, $\gamma'(k) = 0$. $\delta'(k)$ *will be chosen such that* $\mathrm{d}\delta'/\mathrm{d}k \geq 1$ *(so that* $\delta'(k) > \beta'(k)$*) and*

$$\left.\frac{\mathrm{d}}{\mathrm{d}k}\bar{\Delta}'(k)\right|_{k=0} \geq 0. \tag{5.282}$$

This requires that either $1 - \xi_1^2 - \eta_1\eta_2 > 0$ *and*

$$\left.\frac{\mathrm{d}\delta'}{\mathrm{d}k}\right|_{k=0} \geq \frac{1 - \beta(1 - \xi_1^2 - \eta_1\eta_2}{\beta(1 - \xi_1^2 + \eta_1\eta_2)}, \tag{5.283}$$

or $1 - \xi_1^2 - \eta_1\eta_2 = 0$ *and* $(1 - \xi_1^2 - \eta_1\eta_2)\beta \geq 1$*, which implies* $|\eta_1| = |\eta_2| = \sqrt{1 - \xi_1^2}$ *and* $\eta_1\eta_2 < 0$ *which would mean that* $\bar{\Delta}$ *would be independent of* δ*, which could then be increased to increase* $\det Q^{-1}$.

Now inequality (5.283) is a lower bound on $\mathrm{d}\delta'/\mathrm{d}k$ *and so* $\delta'(k)$ *can be chosen to increase sufficiently rapidly and then* $\det Q'(k)^{-1}$ *will also increase.*

If $(\delta-\alpha)(\delta-\beta)-\gamma^2=0$, $\alpha,\beta<\delta$, *then* $\gamma\neq 0$, *and so* $|\gamma|$ *can be decreased to increase* $\det Q^{-1}$ *without affecting* $\bar{\Delta}$.

Hence, if $\mathrm{I}-\delta^{-1}\bar{Q}^{-1}\geq 0$ *holds but* $\mathrm{I}-\delta^{-1}\bar{Q}^{-1}>0$ *does not hold,* $\det Q^{-1}$ *can be increased to yield an ellipsoid with smaller volume.*

Now the weaker condition

$$u^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})u=v^{\mathrm{T}}(\mathrm{I}-\delta^{-1}\bar{Q}^{-1})v=0, \tag{5.284}$$

for linearly independent u,v *will be investigated. If these equations are solved for* α *and* γ, *it will be seen that*

$$\det Q^{-1}=\frac{\delta^{p-2}}{4u_1^2v_1^2}[-(u_1v_2+u_2v_1)^2\delta^2+2(u_1^2v_2^2+2u_1^2v_1^2+u_2^2v_1^2)\beta\delta-(u_1v_2-u_2v_1)^2\beta^2], \tag{5.285}$$

provided that $u_1,v_1\neq 0$ ($u_1v_2\neq u_2v_1$ *as* u *and* v *are linearly independent, so* $u_1=v_1=0$ *is impossible). Set* $\beta'(k)=\beta k$, $\delta'(k)=\delta k$ *and let* $(\alpha'(k),\gamma'(k))$ *be the solution of equations (5.284), with* α, β, γ *and* δ *replaced by* $\alpha'(k)$, $\beta'(k)$, $\gamma'(k)$ *and* $\delta'(k)$. *The quantities* $\alpha'(k)$, $\beta'(k)$ *and* $\delta'(k)$ *will remain positive so long as* k *belongs in a certain neighbourhood, as will* $\alpha'(k)\beta'(k)-\gamma'(k)^2$. *As equation (5.285) becomes homogeneous of degree 2 in* k *if* β *and* δ *are replaced by* $\alpha'(k)$ *and* $\delta'(k)$ *respectively, it increases as* k *increases. In addition,* $\bar{\Delta}'(k)$ *will increase as* $\bar{\Delta}\geq 0$. *If* $u_1=0$ *or* $v_1=0$, $\delta=\beta$ *and* $\alpha=\delta-2v_2\gamma/v_1$. *Here, the assignments* $\alpha'(k)=\alpha$, $\beta'(k)=\delta'(k)=\delta k$, $\gamma'(k)=\gamma k$ *will result in* $\det Q^{-1}$ *increasing and* $\bar{\Delta}'(k)$ *remaining nonnegative.*

Thus, an ellipsoid with smaller volume can also be chosen when the conditions (5.285) hold.

So, if $\xi_0\neq\xi_1$,

$$\begin{aligned}\alpha &= \frac{1+\beta(\eta_1-\eta_0)(\eta_2-\eta_0)-\delta[1-\xi_1^2-\eta_1\eta_2-(\eta_1+\eta_2)\eta_0]}{(\xi_1-\xi_0)^2}\\ \gamma &= \frac{\delta(\eta_1+\eta_2)-\beta(\eta_1+\eta_2-2\eta_0)}{2(\xi_1-\xi_0)}.\end{aligned} \tag{5.286}$$

These lead to

$$\begin{aligned}\det Q^{-1} &= \delta^{p-2}(\alpha\beta-\gamma^2)\\ &= \frac{4\beta-(\eta_2-\eta_1)^2\beta^2+2(2-\eta_1^2-\eta_2^2-2\xi_1^2)\beta\delta+(\eta_1+\eta_2)^2\delta}{(\xi_1-\xi_0)^2}.\end{aligned} \tag{5.287}$$

If this quantity is not a maximum for the values of β *and* δ *which enter into the definition of* $\mathcal{E}_2$, *then* β *and* δ *can be varied to give a larger value of* $\det Q^{-1}$. *But a necessary condition for a maximum*

of $\det Q^{-1}$ *is that*

$$\begin{aligned}\frac{\partial}{\partial\beta}\det Q^{-1} &= \frac{4-2(\eta_2-\eta_1)^2\beta+2(2-\eta_1^2-\eta_2^2-2\xi_1^2)\delta}{(\xi_1-\xi_0)^2},\\ \frac{\partial}{\partial\delta}\det Q^{-1} &= \frac{2(2-\eta_1^2-\eta_2^2-2\xi_1^2)\beta-2(\eta_1+\eta_2)^2\delta}{(\xi_1-\xi_0)^2}\end{aligned} \tag{5.288}$$

both vanish. If $\eta_1^2, \eta_2^2 \neq 1-\xi_1^2$*, this requires that*

$$\beta = -\frac{(\eta_1+\eta_2)^2}{(1-\xi_1^2-\eta_1^2)(1-\xi_1^2-\eta_2^2)}, \tag{5.289}$$

which is negative, a contradiction. If either $\eta_1^2 = 1-\xi_1^2$ *or* $\eta_2^2 = 1-\xi_1^2$*, the quantities in equations (5.288) cannot both vanish. Hence,* $\det Q^{-1}$ *cannot have a maximum, and so* $\mathcal{E}_2$ *cannot have exactly two contact points.*

If $\xi_0 = \xi_1$*, then equations (5.286) can be satisfied for any* α *and* γ*, so* $\det Q^{-1}$ *can be increased by varying these quantities, and again the possibility that* $\mathcal{E}_2$ *has exactly two contact points can be rejected.* ■

Theorem 5.39:

If the points (ξ_1, η) are Type 1 contact points for $\eta \in (\eta_1, \eta_2) \subset [-\sqrt{1-\xi_1^2}, \sqrt{1-\xi_1^2}]$, where $\eta_1 < \eta_2$, then there is at least one further contact point.

□

Proof 5.39: *This is a simple consequence of the fact that the non-degenerate ellipsoid*

$$\mathcal{E}\left(\begin{bmatrix}\xi_0\\ \eta_0\\ 0\end{bmatrix}, \begin{bmatrix}\frac{1-\delta(1-\xi_1^2-\eta_0^2)}{(\xi_1-\xi_0)^2} & \frac{\delta\eta_0}{\xi_1-\xi_0} & 0\\ \frac{\delta\eta_0}{\xi_1-\xi_0} & \delta & 0\\ 0 & 0 & \delta\mathrm{I}_{p-2}\end{bmatrix}\right) \tag{5.290}$$

can be continuously replaced by a smaller one passing through the same Type 1 points by varying δ*.*

■

Lemma 5.40: Let $\mathcal{E}_2$ be the minimum-volume ellipsoid about $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$. Suppose the set of contact points consists of n_{0b} T0b sets, n_{1a} T1a sets, n_{1b} T1b sets, n_2 T2 sets and $n_{\mathbb{H}}$ $\mathbb{H}$T sets. Then $\mathcal{E}_2$ is also an ellipsoid of locally minimum volume among the family of ellipsoids containing the same T1b, T2 and $\mathbb{H}$T sets, and n_{1a} T1a and n_{0b} sets, where the T1a sets remain associated with their original hyperplanes.

□

Proof 5.40: *Suppose* $\mathcal{E}_2$ *can be continuously varied to give a smaller ellipsoid* $\mathcal{E}_2'$*, where the variation does not change the T1b, T2 and* $\mathbb{H}$*T sets, but may move the points in the T0b and T1a*

sets, while maintaining the tangency and Jacobian conditions for the points in all the sets. Then, there must be a variation $\mathcal{E}_2''$ of $\mathcal{E}_2'$ which does not introduce new contact points but has smaller volume than $\mathcal{E}_2$. But then $\mathcal{E}_2'' \supset \mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ and $\mathrm{vol}\mathcal{E}_2'' < \mathrm{vol}\mathcal{E}_2$, contradicting the minimality of $\mathcal{E}_2$. ■

Here the results of the Theorems will be summed up in some rules:

1. There are at least three contact points (Theorem 5.38);
2. If there is a HT set, then either there is a further HT set, or there is at least one T0b, T1a, T1b or T2 set (Theorem 5.39);
3. If there is a T1a set, no further contact point can lie on its hyperplane (Theorem 5.25);
4. There are no Type 0a points (Theorem 5.23);
5. There is at most one T0b set (Theorem 5.24);
6. There is no T1a set opposite two T1b sets (Theorem 5.27);
7. There is no T0b set if there is an HT set (Theorem 5.28);
8. There cannot be a T1a set opposite an HT set (Theorem 5.27);
9. There cannot be two T1b sets opposite an HT set (Theorem 5.27);
10. If there are two HT sets, there cannot be a T1a set (Theorem 5.29);
11. Theorem 5.30 holds;
12. There are at most four Type ib points (Theorem 5.31).
13. If there is a Type 0b point, there are at most three points Type ib points (Theorem 5.32).
14. If there is a HT set, there are at most two points Type ib points (Theorem 5.33).
15. If there is a HT set and a Type 1a point, there are no Type ib points (Theorem 5.34).

In Figure 5.3 the possible dispositions of the hyperplanes relative to the original ellipsoid are illustrated. The possible combinations of the Types of sets are given in Tables 5.3 to 5.3. The example of the case (3, 1) of Figure 5.3 is illustrative.

There are four possible candidate T1b sets and one candidate T2 set. Each of the three possible HT sets can supply at most one T1a set (rule 3). In what follows, "choosing a T0b

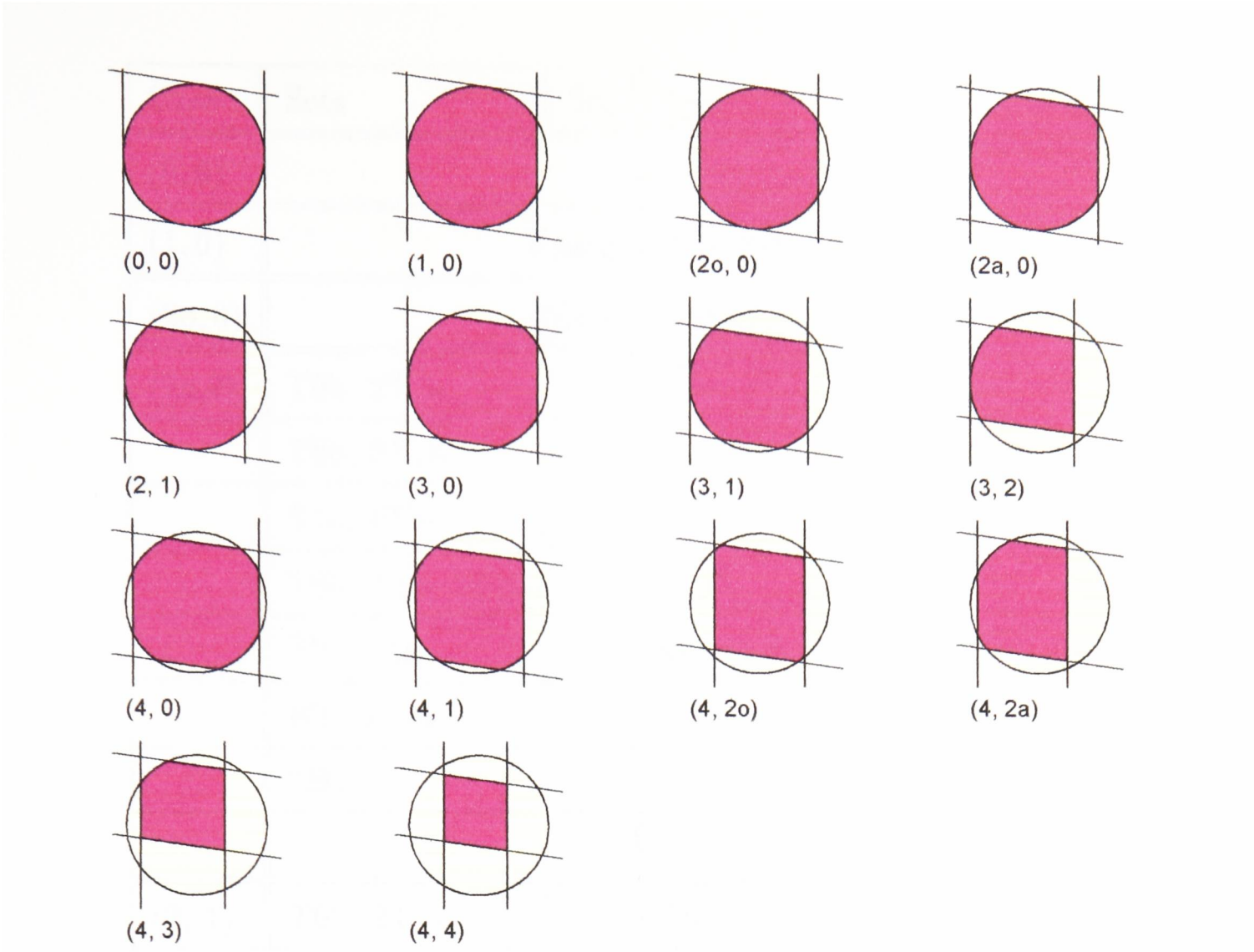


Figure 5.3: The possible cases, ignoring occurrences of Type 2b points. The first digit gives the number of hyperplanes bounding $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, with "o" (for opposite) or "a" (for adjacent) appended where necessary. The second digit gives the number of pairs of hyperplanes which intersect in the interior of $\mathcal{E}_0$, again with "o" or "a" appended as appropriate.

set" means choosing whether there exists a T0b set, and "choosing a T1b set" means choosing the hyperplane on which the Type 1a point lies. If there are three contact points, these might consist of the union of one T0b and two T1a sets (3 ways of choosing these); or one T0b, one T1a and one T1b set (8 — T1b cannot be on same hyperplane as the T1a set — rule 3); or one T0b, one T1a and one T2 set (1 — rule 3); or one T0b and two T1b sets (5 — rule 11); or one T0b, one T1b and one T2 set (4); or three T1a sets (1); or two T1a sets and one T1b (4 — rule 3); or two T1a sets and one T2 (0 — here, only theoretically possible — at least one of the T1a sets' hyperplanes must contain the point in the T2 set); or one T1a and two T1b sets (5 — rule 6); or one T1a, one T1b and one T2 set (2 — rule 3); or three T2b sets (3 — rule 11); or, finally, two T1b and one T2 set (6). Thus, there are 42 ways of choosing three contact points.

When there are four contact points to be chosen, this can be done by choosing one T0b and three T1a sets (1); or one T0b, two T1a sets and one T1b (4); or one T0b, one T1a and two

Case	Sets	No.	Sets	No.	Total
(0,0)	$\mathcal{E}_2 = \mathcal{E}_0$				
(1,0)	König and Pallaschke				
(2o,0)	König and Pallaschke				
(2a,0)	T0b, 2T1a	1	T0b, T1a, T1b	4	
	T0b, 2T1b	4	T1a, 2T1b	2	
	T1a, 2T1b	2	3T1b	4	17
	T0b, T1a, 2T1b	2	4T1b	1	3
	HT, T1a	2	HT, T1b	4	6
	HT, 2T1b	2			2
	2HT	1			1
					29
(2, 1)	T0b, 2T1a	1	T0b, T1a, T1b	2	
	T0b, 2T1b	1	T0b, T1b, T2	2	
	2T1b, T2	1			7
	T0b, 2T1b, T2	1			1
	HT, T1b	2			2
	2HT	1			1
					11

Table 5.1: Number of different combinations of sets of contact points for Cases (0, 0) to (2, 1).

T1b sets (5); or one T0b, one T1a, one T1b and one T2 set (2); or, one T0b, two T1b and one T2 set (6); or one T1a, two T1b and one T2 set (1 — rule 3); or four T1b sets (1); or three T1b and one T2 set (6). So, there are 26 ways of doing this.

When there are five contact points to be chosen, this can be done by choosing one T0b, one T1a, two T1b and one T2 set (1) or four T1b and one T2 set (1). So, there are 2 ways of doing this.

There are no ways of choosing six contact points, but the choice of one HT and one further point can be made by choosing one HT and one T1a (3 — rule 8); or, one HT and one T1b (8); or, one HT and one T2 (1). This can therefore be done in 12 ways.

The choice of one HT and two further points can be made by choosing one HT and two T1b

Case	Sets	No.	Sets	No.	Total
(3, 0)	T0b, 2T1a	3	T0b, T1a, T1b	12	
	T0b, 2T1b	12	3T1a	1	
	2T1a, T1b	6	T1a, 2T1b	16	
	3T1b	16			65
	T0b, 3T1a	1	T0b, 2T1a, T1b	6	
	T0b, T1a, 2T1b	16	T1a, 3T1b	8	
	4T1b	14			45
	$\mathbb{HT}$, T1a	3	$\mathbb{HT}$, T1b	12	15
	$\mathbb{HT}$, 2T1b	16			16
	2$\mathbb{HT}$	3			3
					144

Table 5.2: Number of different combinations of sets of contact points for Case (3, 0).

sets (6 — rule 9); or, one $\mathbb{HT}$, one T1b and one T2 set (2). This can therefore be done in 8 ways.

The choice of one $\mathbb{HT}$ and three further points can be made by choosing one $\mathbb{HT}$, two T1b and one T2 set (1).

Finally, there are three ways of choosing two $\mathbb{HT}$ sets. Any further contact point would have to be a Type 2 point, but the only available such point lies in at least one of the $\mathbb{HT}$ sets. Thus, there are 106 possible ways of choosing the connected components of the set of contact points. For each of the sets of contact points present, there is a polynomial equality to be satisfied, possibly augmented by nonempty tangent and Jacobian conditions expressed as polynomial inequalities. In addition, equation (5.38) holds. In order to find the BLJ ellipsoid about $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, the following procedure can be adopted:

1. if ζ_1, ζ_2 obey the conditions of Theorem 5.12 or Corollary 5.14 as appropriate, then $\mathcal{E}_2 = \mathcal{E}_0$;

2. the disposition of the possible contact points according to the classification of Figure 5.3 is determined;

3. the appropriate Table 5.3 to 5.3 is worked through, from top to bottom. If an ellipsoid containing $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ is found corresponding to an entry in one of the sections of the

Case	Sets	No.	Sets	No.	Total
(3,1)	T0b, 2T1a	3	T0b, T1a, T1b	8	
	T0b, T1a, T2	1	T0b, 2T1b	5	
	T0b, T1b, T2	4	3T1a	1	
	2T1a, T1b	4	T1a, 2T1b	5	
	T1a, T1b, T2	2	3T1b	3	
	2T1b, T2	6			42
	T0b, 3T1a	1	T0b, 2T1a, T1b	4	
	T0b, T1a, 2T1b	5	T0b, T1a, T1b, T2	2	
	T0b, 2T1b, T2	6	T1a, 2T1b, T2	1	
	4T1b	1	3T1b, T2	6	26
	T0b, T1a, 2T1b, T2	1	4T1b, T2	1	2
	$\mathbb{H}$T, T1a	3	$\mathbb{H}$T, T1b	8	
	$\mathbb{H}$T, T2	1			12
	$\mathbb{H}$T, 2T1b	6	$\mathbb{H}$T, T1b, T2	2	8
	$\mathbb{H}$T, 2T1b, T2	1			1
	2$\mathbb{H}$T	1			1
					92

Table 5.3: Number of different combinations of sets of contact points for Case (3, 1).

Table, then the BLJ ellipsoid for $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ will be the minimum-volume ellipsoid containing $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$ from that section, as sections below will correspond to ellipsoids conforming to a greater number of constraints than at least one ellipsoid from the higher sections;

4. for each of the combinations of contact points listed in the Table:

   (a) the appropriate polynomial equations are determined;

   (b) $\det Q^{-1} = \delta^{p-2}(\alpha\beta - \gamma^2)$ is maximised subject to these equations. This will require that further polynomial equations be solved in conjunction with those arising from the proposed existence of the contact points and equation (5.6);

   (c) if the resulting ellipsoid does not contain $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, further ellipsoids may be

Case	Sets	No.	Sets	No.	Total
(3,2)	T0b, 2T1a	3	T0b, T1a, T1b	4	
	T0b, T1a, T2	2	3T1a	1	
	2T1a, T1b	4	2T1a, T2	2	
	T1a, T1b, T2	2	2T1b, T2	2	
	T1b, 2T2	2			22
	T0b, 3T1a	1	T0b, 2T1a, T1b	4	
	T0b, 2T1a, T2	2	T0b, T1a, T1b, T2	2	
	T0b, 2T1b, T2	2	T0b, T1b, 2T2	2	
	2T1b, 2T2	1			14
	T0b, 2T1b, 2T2	1			1
	$\mathbb{H}$T, T1b	4	$\mathbb{H}$T, T2	2	6
	$\mathbb{H}$T, T1b, T2	2			2
	2$\mathbb{H}$T	3			3
					48

Table 5.4: Number of different combinations of sets of contact points for Case (3, 2).

determined by converting the tangent or Jacobian inequalities to equalities. The resulting collection of polynomials can then be completed by following step 4b above.

Note that the derivation of the polynomials for each combination of contact points needs to be done once only. This will produce a set of polynomials in α, β, γ, δ, ξ_0 and η_0 with coefficients in terms of geometric quantities related to the disposition of the hyperplanes relative to the original ellipsoid. For each particular occurrence of a combination of contact points, the relevant geometric quantities will be substituted into the coefficients to produce the corresponding set of polynomials for that combination.

Although it is not certain that each particular occurrence of a combination in the appropriate Table will result in a valid ellipsoid, it is certain that at least one will result in the BLJ ellipsoid. It will now be useful to give a Lemma enabling the reduction of a set of polynomials to a more tractable set.

Lemma 5.41: Suppose A_1 and A_2 are irreducible polynomials in u_1 and u_2. Let S be the solution set of $A_1(u_1)A_2(u_2) = 0$; i.e., $S = \{(u_1, u_2) : A_1(u_1)A_2(u_2) = 0\}$. Then $S = (\cup_i S_i) \cup (\cup_j S'_j)$,

Case	Sets	No.	Sets	No.	Total
(4, 0)	T0b, 2T1a	6	T0b, T1a, T1b	24	
	T0b, 2T1b	6	3T1a	4	
	2T1a, T1b	36	T1a, 2T1b	56	
	3T1b	12			144
	T0b, 3T1a	4	T0b, 2T1a, T1b	36	
	T0b, T1a, 2T1b	56	4T1a	1	
	3T1a, T1b	8	2T1a, 2T1b	28	
	T1a, 3T1b	20	4T1b	13	153
	T0b, 4T1a	1	T0b, 3T1a, T1b	8	
	T0b, 2T1a, 2T1b	28	2T1a, 3T1b	8	
	T1a, 4T1b	18			63
	$\mathbb{HT}$, T1a	8	$\mathbb{HT}$, T1b	24	32
	$\mathbb{HT}$, 2T1a	4	$\mathbb{HT}$, 2T1b	56	60
	2$\mathbb{HT}$	6			6
					460

Table 5.5: Number of different combinations of sets of contact points for Case (4, 0).

where S_i is the solution set of $B_{1i}(u_1, u_2) = B_{2i}(u_2) = 0$ and S'_j is the solution set of $C_j(u_1, u_2) = 0$ for some polynomials B_{1i} and C_j in u_1 and u_2 and B_{2i} in u_2. □

Proof 5.41: *If A_1 and A_2 are not irreducible, they can be written $A_1 = A_{11}A_{12}\dots A_{1m}$ and $A_2 = A_{21}A_{22}\dots A_{2n}$ for irreducible A_{ij}, and then $S = \cup_{ij} S_{ij}$, where S_{ij} is the solution set of $A_{1i}(u_1, u_2) = A_{2j}(u_1, u_2) = 0$.*

Now assume A_1 and A_2 are irreducible. If $\deg_1 A_2 > 0$ (where $\deg_1$ is the degree in u_1 of a polynomial), there exist unique rational functions F_Q, F_R, polynomial in u_1 which are such that $A_1 = F_Q A_2 + F_R$, where $\deg_1 F_R < \deg_1 A_2$, by the Eulerian property of polynomials of one variable.

If $F_R = 0$, $F_Q \in \mathbb{R}$ by the irreducibility of A_1 and A_2, and A_1 and A_2 can be replaced by $C_1 = A_1$, that is $S = S_1$.

If $F_R \neq 0$, there exist unique polynomials P_Q, P_R, P_{QD}, P_{RD} such that

1. *$F_Q = P_Q/P_{QD}$, $F_R = P_R/P_{RD}$;*

Case	Sets	No.	Sets	No.	Total
(4, 1)	T0b, 2T1a	6	T0b, T1a, T1b	18	
	T0b, T1a, T2	2	T0b, 2T1b	13	
	3T1a	4	2T1a, T2	12	
	2T1a, T2	1	T1a, 2T1b	30	
	T1a, T1b, T2	6	3T1b	18	
	2T1b, T2	15			125
	T0b sets and set families from previous section of table			76	
	4T1a	1	3T1a, T1b	6	
	2T1a, 2T1b	9	2T1a, T1b, T2	2	
	T1a, 3T1b	42	T1a, 2T1b, T2	12	
	4T1b	11	3T1b, T2	20	179
	T0b sets and set families from previous section of table			30	
	2T1a, 2T1b, T2	1	T1a, 4T1b	6	
	4T1b, T2	11			48
	$\mathbb{HT}$, T1a	8	$\mathbb{HT}$, T1b	18	
	$\mathbb{HT}$, T2	2			28
	$\mathbb{HT}$, 2T1a	6	$\mathbb{HT}$,T1a, T2	2	
	$\mathbb{HT}$, 2T1b	48	$\mathbb{HT}$, T1b, T2	4	60
	$\mathbb{HT}$, 2T1b, T2	12			12
	2$\mathbb{HT}$	6			6
	2$\mathbb{HT}$, T2	1			1
					459

Table 5.6: Number of different combinations of sets of contact points for Cases (4, 1). The "set families from previous section of table" are, of course, those that do not already contain a T0b set and contain less than four T1b points.

2. *P_{QD} and P_{RD} are polynomials in u_2 alone and are monic;*

3. *P_Q and P_{QD} are coprime;*

4. *P_R and P_{RD} are coprime;*

Then $P_{QD}P_{RD}A_1 = P_{RD}P_QA_2 + P_{QD}P_R$ and S is contained in the solution set of $A_2(u_1, u_2) =$

Case	Sets	No.	Sets	No.	Total
(4, 2o)	T0b, 2T1a	6	T0b, T1a, T1b	8	
	T0b, T1a, T2	4	T0b, 2T1b	6	
	T0b, T1b, T2	4	T0b, 2T2	1	
	3T1a	4	2T1a, T1b	12	
	2T1a, T2	2	T1a, 2T1b	4	
	T1a, T1b, T2	8	3T1b	4	
	2T1b, T2	12	T1b, 2T2	4	79
	T0b sets and set families from previous section of table			46	
	4T1a	1	3T1a, T1b	4	
	2T1a, 2T1b	6	2T1a, T1b, T2	4	
	T1a, 2T1b, T2	4	4T1b	1	
	3T1b, T2	6	2T1b, 2T2	6	78
	T0b sets and set families from previous section of table			25	
	2T1a, 2T1b, T2	2	T1a, 3T1b, T2	4	
	4T1b, T2	2	3T1b, 2T2	4	37
	T0b sets and set families from previous section of table			2	
	4T1b, 2T2	1			3
	$\mathbb{H}$T, T1a	8	$\mathbb{H}$T, T1b	12	
	$\mathbb{H}$T, T2	4			24
	$\mathbb{H}$T, 2T1a	4	$\mathbb{H}$T, 2T1b	12	
	$\mathbb{H}$T, T1b, T2	12			28
	$\mathbb{H}$T, 2T1b, T2	4			4
	2$\mathbb{H}$T	6			6
	2$\mathbb{H}$T, T2	2			2
					263

Table 5.7: Number of different combinations of sets of contact points for Case (4, 2o).

$P_{QD}(u_2)P_R(u_1, u_2) = 0$.

Suppose $P_{QD}P_R = A_{31} \dots A_{3k}$ *where each* A_{3i} *is irreducible. Then* S *is contained in the union of the solution sets of* $A_2(u_1, u_2) = A_{3i}(u_1, u_2) = 0$, *where, of course, each* A_{3i} *satisfies* $\deg_1 A_{3i} <$

Case	Sets	No.	Sets	No.	Total
(4, 2a)	T0b, 2T1a	6	T0b, T1a, T1b	12	
	T0b, T1a, T2	4	T0b, 2T1b	5	
	T0b, T1b, T2	8	3T1a	4	
	2T1a, T1b	12	2T1a, T2	2	
	T1a, 2T1b	12	T1a, T1b, T2	10	
	T1a, 2T2	1	3T1b	4	
	2T1b, T2	12	T1b, 2T2	4	96
	T0b sets and set families from previous section of table			57	
	4T1a	1	3T1a, T1b	4	
	3T1a, T2	4	2T1a, 2T1b	6	
	2T1a, T1b, T2	2	T1a, 3T1b	4	
	T1a, 2T1b, T2	8	T1a, T1b, 2T2	2	
	4T1b	1	3T1b, T2	8	
	2T1b, 2T2	6			103
	T0b sets and set families from previous section of table			33	
	T1a, 3T1b, T2	2	T1a, 2T1b, 2T2	1	
	4T1b, T2	2	3T1b, 2T2	6	44
	T0b sets and set families from previous section of table			1	
	4T1b, 2T2	1			2

Table 5.8: Number of different combinations of finite sets of contact points for Case (4, 2a).

$\deg_1 A_2$.

Applying the same argument, where $\deg_1 A_{3i} > 0$, *to each of the pairs* A_2 *and* A_{3i}, *leads to the conclusion that the Lemma holds.* ■

Corollary 5.42: Suppose $A_1, \ldots, A_m$ are polynomials in $u_1, \ldots, u_n$. Then the solution set S of $A_1(u_1, \ldots, u_n) = \cdots = A_n(u_1, \ldots, u_n) = 0$ is contained in $\cup_j S_j$, where S_j is the solution set of $B_j(u_2, \ldots, u_n) = C_{j1}(u_1, \ldots, u_n) = \cdots = C_{jL_j}(u_1, \ldots, u_n) = 0$, for some polynomials B_j, C_{jk}, where possibly $L_j = 0$. □

Proof 5.42: *Follows from the Lemma by noting that* u_2 *can stand for the set* $\{u_2, \ldots, u_n\}$ *and, if the union of the solution sets of* $D_i(u_1) = E_i(u_1, u_2) = 0$ *contains the solution set of* $F(u_1, u_2) =$

Case	Sets	No.	Sets	No.	Total
(4, 2a)	Finite sets				245
	HT, T1a	8	HT, T1b	12	
	HT, T2	4			24
	HT, 2T1a	4	HT, T1a, T2	8	
	HT, 2T1b	12	HT, T1b, T2	10	
	HT, 2T2	1			35
	HT, 2T1b, T2	8	HT, T1b, 2T2	2	10
	2HT	6			6
	2HT, T2	2			2
					322

Table 5.9: Number of different combinations of sets of contact points for Case (4, 2a), where at least one set is infinite.

$G(u_1, u_2) = 0$, *then the union of the solution sets of* $D_i(u_1) = E_i(u_1, u_2) = H(u_1, u_2) = 0$ *contains the solution set of* $F(u_1, u_2) = G(u_1, u_2) = H(u_1, u_2) = 0$. ■

The following Corollaries are immediate consequences of the preceding one.

Corollary 5.43: Suppose $A_1, \ldots, A_m$ are polynomials in $u_1, \ldots, u_n$. Then the solution set S of $A_1(u_1, \ldots, u_n) = \cdots = A_n(u_1, \ldots, u_n) = 0$ is contained in $\cup_j S_j$, where S_j is the solution set of $B_j(u_1, \ldots, u_n) = C_{jl}(u_2, \ldots, u_n) = \cdots = C_{jL_j}(u_2, \ldots, u_n) = 0$, for some polynomials B_j, C_{jk}, where possibly $L_j = 0$. □

Corollary 5.44: Suppose $A_1, \ldots, A_m$ are polynomials in $u_1, \ldots, u_n$. Then the solution set S of $A_1(u_1, \ldots, u_n) = \cdots = A_n(u_1, \ldots, u_n) = 0$ is contained in $\cup_j S_j$, where S_j is the solution set of either $B_j(u_1, \ldots, u_n) = C_{jl}(u_2, \ldots, u_n) = C_{j2}(u_3, \ldots, u_n) = \cdots = C_{jL_j}(u_{L_j+1}, \ldots, u_n) = 0$, for some polynomials B_j, C_{jk}, when $L_j > 0$, or of $B_j(u_1, \ldots, u_n) = 0$. □

The above Corollary may enable the reduction of the polynomials to a tractable set, but it does not guarantee that the initial set of polynomials is simultaneously soluble, nor that a solution necessarily consists of an isolated point. It also generates spurious solutions. However, in the current context, at least one set of polynomials has an isolated solution, as the BLJ ellipsoid is guaranteed to exist, and spurious solutions will automatically be eliminated by the fact that they will not correspond to ellipsoids, or to ellipsoids not containing $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$, or to larger ellipsoids than the smallest ellipsoid yielded by this method.

Case	Sets	No.	Sets	No.	Total
(4, 3)	T0b, 2T1a	12	T0b, T1a, T1b	6	
	T0b, T1a, T2	6	T0b, 2T1b	1	
	T0b, T1b, T2	6	T0b, 2T2	6	
	3T1a	4	2T1a,T1b	4	
	2T1a, T2	3	T1a, 2T1b	2	
	T1a, T1b, T2	8	T1a, 2T2	6	
	2T1b, T2	3	T1b, 2T2	6	
	3T2	1			74
	T0b sets and set families from previous section of table			37	
	4T1a	1	3T1a, T1b	2	
	2T1a, 2T1b	1	2T1a, T1b, T2	2	
	2T1b, 2T2	3	T1b, 3T2	2	48
	T0b sets and set families from previous section of table			11	
	2T1b, 3T2	1			12
	T0b, 2T1b, 3T2	1			1
	$\mathbb{HT}$, T1a	8	$\mathbb{HT}$, T1b	6	
	$\mathbb{HT}$, T2	6			20
	$\mathbb{HT}$, 2T1a	4	$\mathbb{HT}$, T1a, T2	6	
	$\mathbb{HT}$, 2T1b	2	$\mathbb{HT}$, T1b, T2	8	
	$\mathbb{HT}$, 2T2	2			22
	$\mathbb{HT}$, 2T1b, T2	2	$\mathbb{HT}$, T1b, 2T2	2	4
	2$\mathbb{HT}$	6			6
	2$\mathbb{HT}$, T2	3			3
					190

Table 5.10: Number of different combinations of sets of contact points for Case (4, 3).

The number of cases to be investigated is quite large, but the number reduces considerably for cases (2a, 0) and (2, 1), which are important in non-linear parameter estimation (see Veres and Norton [24]).

Case	Sets	No.	Sets	No.	Total
(4, 4)	3T1a	4	2T1a, T2	4	
	T1a, 2T2	4	3T2	4	16
	4T1a	1	4T2	1	2
	$\mathbb{HT}$, T1a	8	$\mathbb{HT}$, T2	8	16
	$\mathbb{HT}$, 2T1a	4	$\mathbb{HT}$, T1a, T2	8	
	$\mathbb{HT}$, 2T2	4			16
	2$\mathbb{HT}$	6			6
	2$\mathbb{HT}$, T2	4			4
					60

Table 5.11: Number of different combinations of sets of contact points for Case (4, 4).

5.4 $p = 2$

When $p = 2$, the major differences with the case where $p > 2$ are:

1. δ does not exist, and, consequently, equation (5.6) does not apply;

2. Type 1b points do not exist.

The methods which apply when $p > 2$ can still be applied, but, in line with item 2 above, Tables 5.3 to 5.3 become Tables 5.4 to 5.4.

It will be noticed that, when $p = 2$, the number of cases reduces to about 1/3 of its value when $p > 2$.

Case	Sets	No.	Sets	No.	Total
(0,0)	$\mathcal{E}_2 = \mathcal{E}_0$				
(1,0)	König and Pallaschke				
(2o,0)	König and Pallaschke				
(2a,0)	T0b, 2T1b	4	3T1b	4	8
	4T1b	1			1
	HT, T1b	4			4
	HT, 2T1b	2			2
	2HT	1			1
					16
(2, 1)	T0b, 2T1b	1	T0b, T1b, T2	2	
	2T1b, T2	1			4
	T0b, 2T1b, T2	1			1
	HT, T1b	2			2
	2HT	1			1
					8

Table 5.12: Number of different combinations of sets of contact points for Cases (0, 0) to (2, 1).

Case	Sets	No.	Sets	No.	Total
(3, 0)	T0b, 2T1b	12	3T1b	16	28
	4T1b	14			14
	$\mathbb{H}$T, T1b	12			12
	$\mathbb{H}$T, 2T1b	16			16
	2$\mathbb{H}$T	3			3
					89

Table 5.13: Number of different combinations of sets of contact points for Case (3, 0).

Case	Sets	No.	Sets	No.	Total
(3,1)	T0b, 2T1b	5	T0b, T1b, T2	4	
	3T1b	3	2T1b, T2	6	18
	T0b, 2T1b, T2	6	4T1b	1	
	3T1b, T2	6			13
	4T1b, T2	1			1
	$\mathbb{H}$T, T1b	8	$\mathbb{H}$T, T2	1	9
	$\mathbb{H}$T, 2T1b	6	$\mathbb{H}$T, T1b, T2	2	8
	$\mathbb{H}$T, 2T1b, T2	1			1
	2$\mathbb{H}$T	1			1
					51

Table 5.14: Number of different combinations of sets of contact points for Case (3, 1).

Case	Sets	No.	Sets	No.	Total
(3,2)	2T1b, T2	2	T1b, 2T2	2	4
	T0b, 2T1b, T2	2	T0b, T1b, 2T2	2	
	2T1b, 2T2	1			5
	T0b, 2T1b, 2T2	1			1
	ℍT, T1b	4	ℍT, T2	2	6
	ℍT, T1b, T2	2			2
	2ℍT	3			3
					21

Table 5.15: Number of different combinations of sets of contact points for Case (3, 2).

Case	Sets	No.	Sets	No.	Total
(4, 0)	T0b, 2T1b	6	3T1b	12	18
	4T1b	13			13
	ℍT,T1b	24			24
	ℍT, 2T1b	56			56
	2ℍT	6			6
					117

Table 5.16: Number of different combinations of sets of contact points for Case (4, 0).

Case	Sets	No.	Sets	No.	Total
(4, 1)	T0b,2T1b	13	3T1b	18	
	2T1b, T2	15			46
	T0b sets and set families from previous section of table			13	
	4T1b	11	3T1b, T2	20	44
	4T1b, T2	15			15
	$\mathbb{HT}$, T1b	18	$\mathbb{HT}$, T2	2	20
	$\mathbb{HT}$, 2T1b	48	$\mathbb{HT}$, T1b, T2	4	52
	$\mathbb{HT}$, 2T1b, T2	12			12
	2$\mathbb{HT}$	6			6
	2$\mathbb{HT}$, T2	1			1
					192

Table 5.17: Number of different combinations of sets of contact points for Cases (4, 1). The "set families from previous section of table" are, of course, those that do not already contain a T0b set and contain less than four T1b points.

Case	Sets	No.	Sets	No.	Total
(4, 2o)	T0b, 2T1b	6	T0b, T1b, T2	4	
	T0b,2T2	1	3T1b	4	
	2T1b, T2	12	T1b, 2T2	4	31
	T0b sets and set families from previous section of table			16	
	4T1b	1	3T1b, T2	6	
	2T1b, 2T2	6			29
	T0b sets and set families from previous section of table			6	
	4T1b, T2	2	3T1b, 2T2	4	12
	4T1b, 2T2	1			5
	$\mathbb{H}$T, T1b	12	$\mathbb{H}$T, T2	4	16
	$\mathbb{H}$T, 2T1b	12	$\mathbb{H}$T, T1b, T2	12	24
	$\mathbb{H}$T, 2T1b, T2	4			4
	2$\mathbb{H}$T	6			6
	2$\mathbb{H}$T, T2	2			2
					129

Table 5.18: Number of different combinations of sets of contact points for Case (4, 2o).

Case	Sets	No.	Sets	No.	Total
(4, 2a)	T0b, 2T1b	5	T0b, T1b, T2	8	13
	4T1b	1	3T1b, T2	8	
	2T1b, 2T2	6	3T1b	4	
	2T1b, T2	12	T1b, 2T2	4	33
	T0b sets and set families from previous section of table			6	
	4T1b, T2	2	3T1b, 2T2	6	14
	4T1b, 2T2	1			1
	$\mathbb{H}$T, T1b	12	$\mathbb{H}$T, T2	4	16
	$\mathbb{H}$T, 2T1b	12	$\mathbb{H}$T, T1b, T2	10	
	$\mathbb{H}$T, 2T2	1			23
	$\mathbb{H}$T, 2T1b, T2	8	$\mathbb{H}$T, T1b, 2T2	2	10
	$\mathbb{H}$T, 2T1b, 2T2	1			1
	2$\mathbb{H}$T	6			6
	2$\mathbb{H}$T, T2	2			2
					119

Table 5.19: Number of different combinations of sets of contact points for Case (4, 2a).

Case	Sets	No.	Sets	No.	Total
(4, 3)	T0b, 2T1b	1	T0b, T1b, T2	6	
	T0b, 2T2	6	2T1b, T2	3	
	T1b, 2T2	6	3T2	1	23
	T0b sets and set families from previous section of table			10	
	2T1b, 2T2	3	T1b, 3T2	2	15
	T0b sets and set families from previous section of table			5	
	2T1b, 3T2	1			6
	T0b, 2T1b, 3T2	1			1
	$\mathbb{H}$T, T1b	6	$\mathbb{H}$T, T2	6	12
	$\mathbb{H}$T, 2T1b	2	$\mathbb{H}$T, T1b, T2	8	
	$\mathbb{H}$T, 2T2	2			12
	$\mathbb{H}$T, 2T1b, T2	2	$\mathbb{H}$T, T1b, 2T2	2	4
	2$\mathbb{H}$T	6			6
	2$\mathbb{H}$T, T2	3			3
					82

Table 5.20: Number of different combinations of sets of contact points for Case (4, 3).

Case	Sets	No.	Sets	No.	Total
(4, 4)	3T2	4			4
	4T2	1			1
	$\mathbb{H}$T, T2	8			8
	$\mathbb{H}$T, 2T2	4			4
	2$\mathbb{H}$T	6			6
	2$\mathbb{H}$T, T2	4			4
					27

Table 5.21: Number of different combinations of sets of contact points for Case (4, 4).

5.5 Conclusion

The present chapter contains the elements of a "meta-algorithm" to find the BLJ ellipsoid about $\mathcal{E}_0 \cap \Pi_1 \cap \Pi_2$. Section 5.1 provides the basis for the later sections, as well as equation (5.38). Then Section 5.2 provides sufficient conditions for $\mathcal{E}_2 = \mathcal{E}_0$; if these are met, there is no need to proceed further. Finally, Section 5.3 provides a family of sets of polynomials each of which can be used as constraints in the maximisation of $\det Q^{-1}$ to produce further polynomials augmenting the family. The simultaneous solution of one of these augmented families corresponds to the required BLJ ellipsoid.

In order to turn the meta-algorithm into an actual algorithm, it is necessary to reduce the sets of polynomials as far as possible. The result will be a family of sets of polynomials whose coefficients are functions of geometric quantities related to the disposition of the strips relative to the original ellipsoid. The reduction needs to be done once only, and then the members of the family need to be solved with the appropriate geometric quantities for particular cases substituted into their coefficients.

Chapter 6

Conclusions and Further Work

6.1 Conclusions

As well as considering the performance of the Fogel-Huang algorithm both empirically and theoretically, this thesis has looked at the possibility of improving the accuracy of the Fogel-Huang algorithm. As there already exist algorithms for the calculation of the unique minimum-volume ellipsoid containing a given bounded convex set, it would be pointless to provide an algorithm with computational complexity approaching that of those algorithms calculating the minimum-volume ellipsoid.

The first approach considered, that of recycling the data (as contemplated by Belforte and related by Pronzato and Walter[22]), results in average improvements which are considerable after the first few cycles, but provides very little improvement thereafter. The ellipsoid obtained after these cycles is still some way from having the BLJ ellipsoid's volume. Moreover, the most dramatic improvement is for data which are particularly resistant to the Fogel-Huang algorithm, leaving the ellipsoids resulting from more tractable data little changed (i.e., on average, having a characteristic length roughly twice that of the BLJ ellipsoid). Of course, it is useful to have an approach which addresses the particular weaknesses of the Fogel-Huang method, but it would also be nice to have an approach which improves on that method for more general data sets. Recycling also requires that at least some of the data be collected beforehand, so it is a departure from strict "onlineness".

An approach better than that of recycling "blind" is to calculate the volume resulting from the application of the algorithm before applying it, and then choosing the hyperplane pair yielding the smallest volume. This choice of "best" hyperplane pair from those available at each stage,

results in an algorithm which, on average, outperforms both the single pass (through the data) Fogel-Huang algorithm and the recycling version, in that after a number of steps equal to that in a cycle, M, say, the "best" choice algorithm gives an ellipsoid with much smaller characteristic length than the single pass, and, after each additional M steps the ellipsoid has a smaller characteristic length than the ellipsoid from recycling. The "best" hyperplane pair algorithm almost always reaches a smaller ellipsoid after, say, K cycles, and approaches it more rapidly. Of course, this approach is as distant from pure "onlineness" as is recycling.

The next approach to the problem was to consider what happens when two hyperplane pairs "added" to the ellipsoid in the same way that one is "added" in the derivation of the Fogel-Huang algorithm (i.e., the inequalities defining the hyperplane pairs (as degenerate ellipsoids) are added to the inequality defining the ellipsoid , after multiplying by parameters q_1 and q_2). The family of ellipsoids resulting from this "addition" is investigated in order to find the minimum-volume member of the family. Although this is, in general, not the BLJ ellipsoid about the intersection of the two strips bounded by the hyperplane pairs and the original ellipsoid, it was hoped that it would be smaller than the ellipsoids produced by applying the Fogel-Huang algorithm sequentially to one hyperplane pair and the original ellipsoid, and then to the other hyperplane pair and the resultant ellipsoid.

This hope is in fact well-founded, but the resulting ellipsoid, after 12 steps, is still quite far from the BLJ ellipsoid.

Improvements were made in two ways:

1. the two hyperplane pairs were chosen out of strict sequence, so that the first pair was applied in the unmodified Fogel-Huang algorithm, then the first and second were applied in the two-hyperplane modified algorithm, then the first and third, the second and fourth, the third and fifth, and so on, the idea being that information "lost in the approximation" from the pair being re-used would be refreshed;

2. the hyperplanes were kept as they were encountered, and the "best", in the sense of most rapidly diminishing the volume of the resultant ellipsoid as its q-parameter was increased from zero, was selected to be the partner of the current pair in the two-hyperplane algorithm.

The first of these changes results in a slight improvement, and the second, precisely because of its departure from onlineness, leads to a somewhat greater one.

Finally, a change in the basic algorithm was made to produce an s-hyperplane algorithm (where s is not greater than the dimension of the parameter space), which relied on approximating $s-1$ hyperplane pairs by $s-1$ hyperplane pairs which are mutually Q-orthogonal and each Q-orthogonal to the remaining hyperplane pair. The performance of this algorithm is a little disappointing, being between that of the two-hyperplane pair algorithm with the different orders for the application of their hyperplanes. However, for low-dimensional parameter spaces at least, this s-hyperplane pair algorithm appears to be relatively cheap, and might be employed in preference to the two-hyperplane algorithms.

To sum up, algorithms making substantial, if not startling, improvements over Fogel-Huang have been derived, and these might be used where there is room for more expensive methods, but not enough resources to calculate the minimum-volume ellipsoid or to utilise exact descriptions of the feasible parameter set. In particular, where individual parameters are important, such as when there is a simple description of the physically possible parameters consistent with the model (e.g. positive decay rates for processes which do not admit negative ones), use might be made of these algorithms. In addition, where the volume of the feasible parameter set is important, the effective improvement is greater.

These improvements have been the motivation for investigating methods leading to the acutal BLJ ellipsoid about the intersection of two strips and an ellipsoid. When this is derived, it can be used in the same way as the "family-optimal" ellipsoids described above.

However, as the number of cases to be investigated in conjunction with some dispositions of the strips is quite large, "hybrid" methods which use the "family-optimal" ellipsoids when the derivation of the BLJ ellipsoid is too complicated may be useful.

A spin-off from the work of finding the BLJ ellipsoid for the intersection of an ellipsoid and two strips will be the BLJ ellipsoid for the intersection of two half-spaces (in general, with nonparallel boundaries), which corresponds to the cases (2a, 0) and (2, 1) of Figure 5.3 and is useful in non-linear estimation.

6.2 Further Work — Extensions

It would be nice to replace Conjecture 4.12 (which states that the smallest volume is attained by a particular ellipsoid in the family there considered, in the derivation of the s-hyperplane algorithm) by a Proposition!

In conjunction with this, it might possible to improve the procedure employed to find the Q-

orthonormal hyperplane pairs, without increasing the complexity too much, bearing in mind that, when $s = p$, $p(p-1)$ possible shifts of hyperplanes are already considered.

Another possibility for improving on the two- and s-hyperplane pair algorithms is to look for different families of ellipsoids containing the intersection of an ellipsoid and two or more strips bounded by hyperplanes, and then optimising for volume over these families (the problem is finding a family for which the algebra is tractable).

A further extension would be to investigate the expected behaviour of the Fogel-Huang algorithm in more depth, possibly with a simpler model of the process whose parameters are being identified, such as a MIMO system with known inputs, or with inputs with known statistics. Under this heading, experiment design for the algorithm could be included: what sequence of inputs is the most likely to provide information valuable in the context of parameter bounding?

A track to investigate further is the possibility of recycling the modified algorithms. The first results show that the improvement due to choosing the hyperplane pair which leads to the greatest reduction in volume at each step in the single-hyperplane pair Fogel-Huang algorithm is far greater than that due to utilising the modified algorithms in fixed-order recycling.

6.3 Further Work — Finding and Exploring the Use of the BLJ Ellipsoid for Two Strips and an Ellipsoid

The major outstanding work is the investigation of the tranformation of the "meta-algorithm" into an algorithm. This require the one-off reduction of the families of sets of polynomials of Chapter 5. Preliminary efforts have been made in this direction, but these have foundered on the limitations on the size of algebraic objects in Maple©[1]. These efforts have used the reduction methods of Chapter 5 in interactive Maple© sessions. It is possible to circumvent the restrictions mentioned by replacing polynomial algebra by array arithmetic (see, e.g., the Appendix), and a great saving in (human) time expenditure will by made by utilising Maple©'s programming facilities rather than working interactively. However, it might be more fruitful to investigate the automated use of Gröbner bases[3] to simultaneously solve the polynomials involved. (Maple©'s built-in Gröbner methods also run into problems concerning the size of the algebraic objects involved here, but it may be possible to avoid this problem).

A work-plan for producting an algorithm for the full one-ellipsoid, two-strip BLJ problem would

[1] $\approx$50 000 terms for a polynomial

first aim at solving cases (2a, 0) and (2, 1), as these are useful elsewhere.
The next topic to be addressed on the solution of (part of) the one-ellipsoid, two-strip BLJ problem would be the investigation of simulations using the resulting ellipsoids, either alone, or in hybrid algorithms with the family-optimal ellipsoids discussed earlier.

6.4 Further Work — Different Directions

Walter and Pronzato[26] mention the possibility of investigating the image of the posterior set S under a transformation T_s. In particular, T_s might map to the characteristic polynomial of the system in question, so $T_{j\omega}(\theta)$ could be examined for stability for each $\theta \in S$. There would be an associated set-valued function, $\mathcal{P}_T$, say, consisting of all possible poles of polynomials in $T_{j\omega}(S)$.
This poses the possibility of using, in ellipsoidal bounding, size criteria which are tailored to a particular map T_s. If the posterior set is reduced to a point, then $\mathcal{P}_T(S)$ would be a set of points and have volume zero. If S were an ellipsoid, what would be the volume of $\mathcal{P}_T(S)$ (not necessarily in the usual measure)? Could this volume be used as a size criterion for ellipsoids in the parameter space?
The errors-in-variables approaches mentioned in Veres and Norton[24] can be adapted to deal with models nonlinear in parameters, as described Pronzato and Walter[22] (the model $y_m(k,\theta) = \Phi(k,\theta)$ is replaced by $y_m(k,\theta) = n_m(k,\theta)^{\mathrm{T}}\theta$ and the regressor error $\epsilon_n(k,\theta) = n_k - n_m(k,\theta)$ is defined, where n_k is known. If it can be said that $e^m_{n_i}(k) \leq \epsilon_{n_i}(k,\theta) \leq e^m_{n_i}(k)$, a transformation from a nonlinear problem to an errors-in-variables one has been made), and then, corresponding to each data item there is a pair of hyperplanes (in general, nonparallel) for each orthant, between which the intersection of the feasible set with the orthant must lie. Clearly, the origin and the orientation of the parameter space must affect any algorithm attempting to utilise this fact to bound the feasible set. The question is, can shifts in the origin and rotations (linear reparameterisations) be used to good effect in this bounding?
Also, attempts should be made to answer the question of which size criteria for ellipsoidal bounding are the most appropriate to what problems.
When the noise bounds for a problem are difficult to estimate in an *a priori* fashion, they can be subsumed into the parameter estimation as additional parameters. For a linear model, the bounds resulting from this approach do not lead to a bounded polytope, but rather to a cone in parameter space. What outer-bounding geometries are analogous in this situation to ellipsoidal

bounds in the case of bounded polytopes?

Finally, if there are rival hypotheses about model structure, can bounding be used to distinguish between them?

Appendix A

Maple© Session for the Modified F-H Algorithm

```
> restart:
> with(linalg):
Warning, new definition for norm
Warning, new definition for trace
> d:= 1 + g1 * q1 + g2 * q2 +(g1 * g2 - h^2) * q1 * q2;
```

$$d := 1 + g1\,q1 + g2\,q2 + (g1\,g2 - h^2)\,q1\,q2$$

```
> n:= (1 + q1 + q2) * d - nu1^2 * q1 * (1 + g2 * q2) + 2 * nu1 * nu2 * h * q1 * q2 - nu2^2 * q2 * (1
  + g1 * q1);
```

$$n := (1 + q1 + q2)\,(1 + g1\,q1 + g2\,q2 + (g1\,g2 - h^2)\,q1\,q2) - \nu1^2\,q1\,(1 + g2\,q2) + 2\,\nu1\,\nu2\,h\,q1\,q2 - \nu2^2\,q2\,(1 + g1\,q1)$$

```
> Delta0:= n^p/d^(p + 1):
> eq1:= numer(factor(diff(ln(Delta0), q1))):
> eq2:= numer(factor(diff(ln(Delta0), q2))):
```

These quantities should be zero to maximise Delta0, i.e., to minimise the ellipsoid volume.
First, collect them wrt to q1:

```
> eq1:= collect(eq1, q1, factor);
```

$$eq1 := -(g1 + q2\,g1\,g2 - q2\,h^2)\,(-1 + p)\,q1^2 - (g1 + q2\,g1\,g2 - q2\,h^2)$$
$$(-q2^2\,g1\,g2 + q2^2\,h^2 - q2\,g1\,g2 - q2\,g1 + \nu2^2\,q2\,g1 - g2\,q2 + 2\,p\,g2\,q2 + \nu1^2\,g2\,q2 + q2\,h^2 - 2\,\nu1\,\nu2\,h\,q2 - g1 - 1 + 2\,p + \nu1^2)\,q1$$
$$+ p\,q2^2\,h^2\,\nu2^2 + g1 - p - \nu2^2\,q2\,g1 - p\,g2^2\,q2^2 + 2\,q2\,g1\,g2 + 2\,q2^2\,g1\,g2 - q2\,h^2 + q2^2\,g1\,g2^2 - q2^2\,h^2\,g2 - \nu2^2\,q2^2\,g1\,g2 + 2\,p\,\nu1^2\,g2\,q2$$
$$+ q2^3\,g1\,g2^2 - q2^3\,h^2\,g2 - 2\,p\,g2\,q2 - 2\,p\,\nu1\,\nu2\,h\,q2 + p\,\nu1^2\,g2^2\,q2^2 + q2^2\,h^2\,\nu2^2 - q2^2\,h^2 + q2\,g1 - 2\,p\,\nu1\,\nu2\,h\,q2^2\,g2 + p\,\nu1^2$$

```
> eq2:= collect(eq2, q1, factor);
```

$$eq2 := (g1\,g2 - h^2)\,(g1 + q2\,g1\,g2 - q2\,h^2)\,q1^3 + (2\,g1\,g2 - g1\,h^2 + q2\,g1^2\,g2^2 - 2\,q2\,g1\,g2\,h^2 - \nu1^2\,g1\,g2 + 2\,q2\,g1\,g2^2 - 2\,g2\,q2\,h^2 + \nu1^2\,h^2$$
$$+ q2^2\,h^4 + q2\,g1^2\,g2 + q2^2\,g1^2\,g2^2 - g1\,q2\,h^2 - p\,q2^2\,h^4 - 2\,q2^2\,g1\,g2\,h^2 - 2\,p\,g1^2\,q2\,g2 + 2\,p\,g1\,q2\,h^2 - p\,q2^2\,g1^2\,g2^2 + 2\,p\,q2^2\,g1\,g2\,h^2 - p\,g1^2$$
$$+ g1^2\,g2 + q2\,h^4 - 2\,p\,\nu1\,\nu2\,h\,g1 - \nu1^2\,g2^2\,q2\,g1 + \nu1^2\,g2\,q2\,h^2 + 2\,\nu1\,\nu2\,h\,q2\,g1\,g2 - 2\,\nu1\,\nu2\,h^3\,q2 - \nu2^2\,g1^2\,q2\,g2 + \nu2^2\,g1\,q2\,h^2 + p\,\nu2^2\,g1^2$$
$$+ p\,h^2\,\nu1^2 - h^2)\,q1^2 + (-2\,q2^2\,h^2\,g2 + 2\,p\,q2^2\,h^2\,g2 + 2\,g1\,g2 + 2\,p\,\nu2^2\,g1 + 2\,q2^2\,g1\,g2^2 + g2 - h^2 + 2\,\nu1\,\nu2\,h\,g2\,q2 - q2\,h^2 + 2\,q2\,g1\,g2^2$$
$$- 2\,p\,q2^2\,g1\,g2^2 - 2\,p\,g1 + g2^2\,q2 - g2\,\nu1^2 + 2\,p\,q2\,h^2 - 2\,g2\,q2\,h^2 - 2\,\nu2^2\,q2\,g1\,g2 + \nu2^2\,q2\,h^2 - \nu1^2\,g2^2\,q2 - 4\,p\,q2\,g1\,g2 - 2\,p\,\nu1\,\nu2\,h$$
$$+ 2\,q2\,g1\,g2)\,q1 + g2 - p\,g2^2\,q2^2 - \nu2^2\,g2\,q2 + p\,\nu2^2 - p - 2\,p\,g2\,q2 + g2^2\,q2^2 + g2^2\,q2 + g2\,q2$$

These are a quadratic and a cubic in q1, respectively.

It is necessary to reduce these polynomials. A first step is to find the leading coefficients (using lcoeff) and to remove any common factors:

```
> l1:= lcoeff(eq1, q1): l2:= lcoeff(eq2, q1): lgcd:= factor(gcd(l1, l2));
```

$$lgcd := g1 + q2\,g1\,g2 - q2\,h^2$$

```
> l1:= factor(l1/lgcd): l2:= factor(l2/lgcd):
```

(Note that lgcd vanishes only for negative q2.)

The first reduction:

```
> eq3:= factor(l1 * eq2 - l2 * q1 * eq1): whattype(eq3), nops(eq3);
```

$$+, 177$$

```
> eq3:= collect(eq3, q1): coeffs(eq3, q1, 't'): t;
```

$$q1^2, q1, 1$$

(As required — coeffs(x, y, 't') puts the powers of y in x into t.)

```
> l1:= lcoeff(eq1, q1): l3:= lcoeff(eq3, q1): lgcd:= factor(gcd(l1, l3));
```

$$lgcd := g1 + q2\,g1\,g2 - q2\,h^2$$

```
> l1:= factor(l1/lgcd): l3:= factor(l3/lgcd):
> eq4:= factor(l1 * eq3 - l3 * eq1): whattype(eq4), nops(eq4);
```

$$*, 2$$

Remove nonzero factors:

```
> op(1, eq4);
```

$$p$$

Number of operands.

```
> nops(op(2, eq4));
```

$$738$$

```
> eq4:= collect(op(2, eq4), q1): coeffs(eq4, q1, 't'): t;
```

$$q1, 1$$

```
> sol1:= q1 = solve(eq4, q1):
```

(sol1 is an equation giving q1 in terms of q2.)

Substitute sol1 into eq1 (this has to be done "piecewise" to avoid algebraic objects which are too large, so sol1 is first used to substitute for q1^2 in terms of q1 and q2).

```
> eq5:= numer(factor(subs(q1^2 = q1 * rhs(sol1), eq1))): whattype(eq5), nops(eq5);
```

+, 3128

```
> nops(expand(denom(rhs(sol1)))), nops(expand(coeff(eq5, q1, 0))), nops(expand(numer(rhs(sol1)))),
  nops(expand(coeff(eq5, q1)));
```

361, 1510, 377, 1618

The size of these objects means that further measures have to be taken to avoid even larger objects intermediate in the next stage of the substitution:

```
> eq6:= 0:
> for i from 1 to 19 do
    tmp:= sum('op(j, denom(rhs(sol1)))', 'j' = 19 * i - 18..19 * i):
    eq6:= eq6 + expand(tmp * coeff(eq5, q1, 0)):
    tmp:= sum('op(j, numer(rhs(sol1)))', 'j' = 19 * i - 18..19 * i):
    eq6:= eq6 + expand(tmp * coeff(eq5, q1)):
  od:
  eq6:= eq6 + expand(sum('op(j, numer(rhs(sol1)))', 'j' = 362..377) * coeff(eq5, q1)): eq6:=
  factor(eq6): whattype(eq6), nops(eq6);
```

*, 6

```
> op(1, eq6) * op(2, eq6) * op(3, eq6) * op(4, eq6) * op(5, eq6);
```

$$-(p+1)(-1+p)^2(g1+q2\,g1\,g2-q2\,h^2)(\nu1+\nu1\,g2\,q2-h\,\nu2\,q2)^3$$

The fourth of these factors cannot vanish for +ve q2, and if the last vanishes:

```
> sol11:= simplify(subs(q2 = nu1/(h * nu2 - g2 * nu1), sol1));
```

$$sol11 := q1 = \frac{\nu2}{h\,\nu1 - g1\,\nu2}$$

But it can easily be shown that q1 = nu2/(h nu1 - g1 nu2), q2 = nu1/(h nu2 - g2 nu1) lead to nonvalid values for the components of the shape matrix of the putative family-optimal ellipsoid.

```
> eq6:= op(6, eq6):
> tmp:= collect(eq6, q2): coeffs(tmp, q2, 't'): t;
```

$$q2, q2^5, q2^3, q2^2, q2^4, 1$$

eq6 is a 5th degree polynomial in q2 one of whose solutions leads to the family-optimal ellipsoid when it is used in conjunction with the equation sol1 for q1.

The coefficients in this polynomial are used in the modified F-H algorithm.

Bibliography

[1] G. Belforte, B. Bona, and V. Cerone. Parameter estimation algorithms for a set-membership description of uncertainty. *Automatica*, 26:887–898, 1990.

[2] M. Berger. *Geometry*, volume 1. Springer-Verlag, Berlin, Heidelberg and New York, 1987.

[3] Bruno Buchberger and Franz Winkler, editors. *Gröbner Bases and Applications.* London Mathematical Society. Cambridge University Press, Cambridge, 1998.

[4] J. C. Burkill and H. Burkill. *A Second Course in Mathematical Analysis.* Cambridge University Press, Cambridge, 1970.

[5] B. C. Carlson. *Special Functions of Applied Mathematics.* Academic Press, New York, 1977.

[6] Cécile Durieu, Boris T. Polyak, and Éric Walter. Ellipsoidal state outer-bounding for MIMO systems via analytical techniques. In *Symposium on Modelling Analysis and Simulation (CESA'96 IMACS Multiconference), Volume 2*, pages 843–848, 1996.

[7] Cécile Durieu, Boris T. Polyak, and Éric Walter. Trace versus determinant in ellipsoidal outer bounding with application to state bounding. In *IFAC'96 World Congress, San Francisco, Volume I*, pages 43–48, 1996.

[8] T. F. Filippova, A. B. Kurzhanski, K.Sugimoto, and I. Vályi. Ellipsoidal state estimation for uncertain dynamical systems. In M. Milanese, J. P. Norton, H. Piet-Lahanier, and Éric Walter, editors, *Bounding Approaches to System Identification*, pages 213–238. Plenum Press, New York, 1996.

[9] Casper Goffman. *Real Functions.* Rinehart, New York (in U.K., Constable, London), 1953.

[10] Herbert Goldstein. *Classical Mechanics.* Addison-Wesley, Reading, Mass., second edition, 1980.

[11] Mohinder S. Grewal and Angus P. Andrews. *Kalman Filtering, Theory and Practice.* Information and System Sciences. Prentice-Hall, New Jersey, 1993.

[12] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometrical Algorithms and Combinatorial Optimization.* Springer-Verlag, Berlin, Heidelberg and New York, 1988.

[13] Luc Jaulin and Eric Walter. Set inversion via interval analysis for nonlinear bounded-error estimation. *Automatica*, 29(4):1053–1064, 1993.

[14] Luc Jaulin and Eric Walter. Guaranteed robust nonlinear parameter bounding. In *Symposium on Modelling Analysis and Simulation (CESA'96 IMACS Multiconference), Volume 2*, pages 1156–1161, 1996.

[15] J.P.Norton. *An Introduction to Identification.* Academic Press, London and New York, 1986.

[16] Hermann König and Diethard Pallaschke. On Khachian's algorithm and minimal ellipsoids. *Numerische Mathematik*, 36:211–223, 1981.

[17] E. K. Kostousova and A. B. Kurzhanski. Theoretical framework and approximation techniques for parallel computation in set-membership state estimation. In *Symposium on Modelling Analysis and Simulation (CESA'96 IMACS Multiconference), Volume 2*, pages 849–854, 1996.

[18] A. B. Kurzhanski and M. Tanaka. On a unified framework for deterministic and stochastic treatment of identification problems. Working paper (wp-89-013), International Institute for Applied Systems Analysis, January 1989.

[19] D. Maksarov and J. P. Norton. Tuning of noise bounds in state bounding. In *Symposium on Modelling Analysis and Simulation (CESA'96 IMACS Multiconference), Volume 2*, pages 837–842, 1996.

[20] O.Mangasarian. *Nonlinear Programming.* McGraw–Hill, New York, 1969.

[21] Luc Pronzato and Éric Walter. Experiment design in a bounded-error context: comparison with D-optimality. *Automatica*, 25(3):383–91, 1989.

[22] Luc Pronzato and Éric Walter. Minimum-volume ellipsoids. *International Journal of Adaptive Control and Signal Processing*, 8:15–30, 1994.

[23] Luc Pronzato and Éric Walter. Volume-optimal inner and outer ellipsoids. In M. Milanese, J. P. Norton, H. Piet-Lahanier, and Éric Walter, editors, *Bounding Approaches to System Identification*, pages 119–138. Plenum Press, New York and London, 1996.

[24] Sàndor M. Veres and John P. Norton. Parameter-bounding algorithms for linear errors in variables models. In *9th IFAC/IFORS Symposium on Identification and Estimation*, pages 1038–1043, 1991.

[25] Éric Walter and Hélène Piet-Lahanier. Estimation of parameter bounds from bounded-error data: a survey. *Special Issue of Mathematics and Computers in Simulation*, 32:449–68, 1990.

[26] Éric Walter and Luc Pronzato. Characterising sets defined by inequalities. In *SYSID'94 Conference, Copenhagen*, pages 15–26, 1994.